ΠΜΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Ανασκόπηση και συγκριτική αξιολόγηση στατιστικών μεθόδων ανάλυσης πληθυσμιακών ερευνών με διαφορετικό δειγματοληπτικό κλάσμα ανά στρώμα παρουσία μη ανταπόκρισης

ΜΑΡΓΕΤΑΚΗ ΑΙΚΑΤΕΡΙΝΗ

Αθήνα, 2015

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών.

Εγκρίθηκε την _____ από την εξεταστική επιτροπή:

ΟΝ/ΜΟ	**ΒΑΘΜΙΔΑ**	**ΥΠΟΓΡΑΦΗ**
Γ. Τουλούμη (Επιβλέπουσα)	Αν. Καθηγήτρια	_____
Ν. Πανταζής	ΕΔΙΠ	_____
Α. Μπουρνέτας	Καθηγητής	_____

Στην αγαπημένη μου αδερφή, Ιωάννα.

Ευχαριστίες

Ευχαριστώ θερμά την επιβλέπουσα καθηγήτρια Γ.Τουλούμη για το εξαιρετικά ενδιαφέρον θέμα που μου ανέθεσε και την κυρία Γ.Βουρλή για την καθοριστική της βοήθεια σε κάθε βήμα μου, καθώς και όλους τους διδάσκοντες του ΠΜΣ Βιοστατιστική ο οποίοι μου μετέδωσαν τις γνώσεις τους και την αγάπη τους προς το αντικείμενο της Βιοστατιστικής.

Περιεχόμενα

Εισαγωγή

Οι μελέτες σε πληθυσμιακό επίπεδο είναι ένα πολύ χρήσιμο και συχνά χρησιμοποιούμενο εργαλείο για την μελέτη των διάφορων χαρακτηριστικών των πληθυσμών. Έχουν ευρεία εφαρμογή σε πολλούς τομείς, στην αποτύπωση της υγείας του πληθυσμού, στα οικονομικά, στην πολιτική κ.α. Ιδιαίτερα, η μελέτη των χαρακτηριστικών που αφορούν στην υγεία, οι επιδημιολογικές μελέτες, είναι εξαιρετικά χρήσιμες στον προγραμματισμό των υπηρεσιών υγείας μπορούν όμως και να αναδείξουν ενδιαφέρουσες συσχετίσεις μεταξύ εκθέσεων σε παράγοντες κινδύνου καικατά αιτίες νοσηρότητα. Οι μελέτες αυτές βασίζονται σε δείγμα το οποίο προέρχεται από τον υπό μελέτη πληθυσμό.

Κατά τη διεξαγωγή μιας τέτοιας μελέτης ιδιαίτερα σημαντικό ρόλο διαδραματίζει ο δειγματοληπτικός σχεδιασμός. Διαδεδομένες μέθοδοι δειγματοληψίας περιλαμβάνουν την απλή τυχαία δειγματοληψία (simple ramdom sampling) , την συστηματική δειγματοληψία (systematic sampling) , την στρωματοποιημένη δειγματοληψία (stratified sampling), την πολυσταδιακή ή κατά συστάδες δειγματοληψία (cluster sampling) καθώς και συνδυασμό των προαναφερθέντων μεθόδων (σύνθετος σχεδιασμός). Τα χαρακτηριστικά της δειγματοληψίας είναι απαραίτητο να λαμβάνονται υπόψιν κατά το στάδιο της στατιστικής ανάλυσης. Συνήθως ο ερευνητής επιλέγει δείγμα με τρόπο ώστε αυτό να είναι αντιπροσωπευτικό του πληθυσμού από τον οποίο προέρχεται. Δηλαδή το κάθε άτομο του πληθυσμού να έχει την ίδια πιθανότητα να συμπεριληφθεί στο δείγμα (equiprobability samping).

Υπάρχουν όμως περιπτώσεις στις οποίες μας ενδιαφέρει να υπερεκπροσωπηθούν συγκεκριμένες σπάνιες ομάδες, ιδιαίτερου ενδιαφέροντος, του πληθυσμού, στο δείγμα μας. Τέτοια μπορεί να είναι η περίπτωση μίας (ίσως και παραπάνω) συγκεκριμένης γεωγραφικής περιοχής με μικρό σχετικά πληθυσμό από την οποία θα επιλέγονταν λίγα άτομα αν εφαρμόζονταν μέθοδοι ίσης πιθανότητας επιλογής. Κάτι τέτοιο θα είχε ως αποτέλεσμα οι σημειακές εκτιμησεις (όπως του επιπολασμού ενός παράγοντα κινδύνου) να συνοδεύονται από μεγάλη αβεβαιότητα, καθιστώντας τις συγκρίσεις των σημειακών εκτιμήσεων μεταξύ γεωγραφικών περιοχών προβληματικές λόγω χαμηλής ισχύος. Στην περίπτωση όμως που επιλεγεί ένας τέτοιος σχεδιασμός το δείγμα παύει να είναι αντιπροσωπευτικό του γενικού πληθυσμού. Αυτό διορθώνεται στο στάδιο της στατιστικής ανάλυσης σταθμίζοντας τις παρατηρήσεις με, ως είναι το συνηθέστερο, το αντίστροφο της πιθανότητας επιλογής στο δείγμα.

Ένα πρόβλημα στο οποίο υπόκεινται οι μελέτες αυτές είναι αυτό της μη-ανταπόκρισης, δηλαδή είτε όταν ένα άτομο το οποίο έχει επιλεγεί στο δείγμα αρνηθεί να συμμετάσχει (μη ανταπόκριση ατόμου - unit nonrespnse), είτε όταν ένα άτομο το οποίο έχει δεχθεί να συμμετάσχει αρνηθεί να απαντήσει κάποιο ή κάποια συγκεκριμένα ερωτήματα (μη ανταπόκριση ερώτησης, item nonrespnse). Είναι σπάνιο έως και ουτοπικό μια μελέτη σε πληθυσμιακό επίπεδο να έχει 100% ποσοστό ανταπόκρισης. Η μη ανταπόκριση μπορεί να εισάγει σφάλμα στις εκτιμήσεις ιδιαίτερα εάν εκείνοι που δεν ανταποκρίνονται διαφέρουν συστηματικά από εκείνους που ανταποκρίνονται ως προς τα χαρακτηριστικά ενδιαφέροντος. Το δείγμα δεν μπορεί να θεωρηθεί ότι διατηρεί τα χαρακτηριστικά τα οποία θα έπρεπε, βάσει σχεδιασμού, να έχει. Αμφιλεγόμενο είναι το ποιο ποσοστό μη-ανταπόκρισης είναι το ελάχιστο ανεκτό. Υπάρχουν διαφορετικές απόψεις επί του θέματος όπως ότι 60% ανταπόκριση είναι καλή και 70% είναι πολύ καλή (Babbie 2002).Άλλη άποψη είναι ότι 85% θεωρείται οριακά αποδεκτό, κάτω από 70% υπάρχει σοβαρή πιθανότητα σφάλματος (Singleton Jr & Bruce n.d.).

Σε κάθε περίπτωση, εκτός του ποσοστού μη ανταπόκρισης, σημαντικό ρόλο διαδραματίζει και ο μηχανισμός παραγωγής ελλειπουσών τιμών.

Προτεινόμενες τεχνικές για την ελαχιστοποίηση της μη ανταπόκρισης περιλαμβάνουν την χρήση απλών και ελκυστικών ερωτηματολογίων, τεχνικές πειθούς από το προσωπικό κ.α. (Dillman et al. 2002, Gary 2007).

Παρά την χρήση τέτοιων τεχνικών ένα ποσοστό μη ανταπόκρισης είναι αναπόφευκτο. Σε αυτές τις περιπτώσεις χρειάζεται να γίνουν κατάλληλες διορθώσεις κατά τη στατιστική ανάλυση οι οποίες όμως θα συνυπολογίζουν και τον σχεδιασμό της μελέτης.

Σκοπός της παρούσας διπλωματικής είναι η ανασκόπηση μεθόδων αντιμετώπισης της μη ανταπόκρισης, σε επίπεδο ατόμου και ερώτησης, σε σύνθετες πληθυσμιακές μελέτες. Στη συνέχεια θα εφαρμοστούν οι προτεινόμενες μέθοδοι και θα αξιολογηθούν, συγκρίνοντας τα αποτελέσματα τους με αποτελέσματα ανάλυσης που αγνοεί την παρουσία μη ανταπόκρισης, στα δεδομένα της Εθνικής Μελέτης Νοσηρότητας και παραγόντων κινδύνου (ΕΜΕΝΟ). Η ΕΜΕΝΟ είναι Πανελλαδική μελέτη καταγραφής των παραγόντων κινδύνου για χρόνια νοσήματα εστιάζοντας κυρίως σε καρδιαγγειακά και αναπνευστικά σε τυχαίο δείγμα του γενικού πληθυσμού ενηλίκων. Το συνολικό δείγμα θα είναι της τάξης των 6000 ατόμων. Για τους σκοπούς της παρούσας διπλωματικής θα χρησιμοποιηθούν τα δεδομένα που έχουν συλλεχθεί κατά την περίοδο Μαρτίου-Δεκεμβρίου.

Μέρος I

Θεωρία

Κεφάλαιο 1

Η E.ΜΕ.ΝΟ

1.1 Γενικά

Η E.ΜΕ.ΝΟ στοχεύει στην παρακολούθηση, καταγραφή και αξιολόγηση της γενικότερης κατάστασης της υγείας του πληθυσμού στην Ελλάδα συμβάλλοντας καθοριστικά στο σχεδιασμό μίας εθνικής στρατηγικής για την υγεία. Συγκεκριμένα η E.ΜΕ.ΝΟ αποσκοπεί στην καταγραφή της νοσηρότητας και των κύριων παραγόντων κινδύνου χρόνιων νοσημάτων, εστιάζοντας κυρίως σε αναπνευστικά και καρδιαγγειακά νοσήματα. Θα παράσχει πολύτιμες πληροφορίες για την αξιολόγηση του βαθμού εφαρμογής των συνιστώμενων μέτρων πρόληψης και των πιθανών φραγμών στην πρόσβαση στο σύστημα υγείας καθώς και των κοινωνικό-οικονομικών παραγόντων που επηρεάζουν την υγεία, και τις επιπτώσεις της οικονομικής κρίσης, σε αυτή. Επίσης, θα επιχειρήσει να χαρτογραφήσει τα επίπεδα ατμοσφαιρικής ρύπανσης και να διερευνήσει τις επιπτώσεις της έκθεσης σε αυτή, στην υγεία των πολιτών.

Τα καρδιαγγειακά νοσήματα αποτελούν την πρώτη αιτία θανάτου στην Ευρώπη (42% στους άνδρες, 52% στις γυναίκες) (*Collaborating Centre for Surveillance of Cardiovascular Diseases. Global Cardiovascular Infobase.* 2010). Στην Ελλάδα, η αλλαγή του τρόπου ζωής των Ελλήνων και η επιδείνωση αναγνωρισμένων παραγόντων κινδύνου, όπως η παχυσαρκία, η καθιστική ζωή και το κάπνισμα, η υπέρταση, η υπερχοληστερολαιμία και ο διαβήτης, έχουν οδηγήσει σε αύξηση των καρδιαγγειακών νοσημάτων. Η Χρόνια Αποφρακτική Πνευμονοπάθεια (ΧΑΠ) και το βρογχικό άσθμα περιλαμβάνονται στα σημαντικότερα αίτια νοσηρότητας και θνησιμότητας παγκοσμίως, συμβάλλοντας σημαντικά στην αύξηση των ιατροφαρμακευτικών δαπανών κάθε κράτους.

Η ανεργία ή εργασιακή επισφάλεια, οι συνθήκες ζωής κατά τα πρώιμα χρόνια, ο εθισμός στον καπνό, στην αιθανόλη και άλλες εξαρτησιογόνες ουσίες, το καθημερινό άγχος, τα διατροφικά προβλήματα, η φτώχεια, η διάκριση, ο κοινωνικός αποκλεισμός και η έλλειψη δικτύων υποστήριξης αναφέρονται από τον Παγκόσμιο Οργανισμό Υγείας μεταξύ των οικονομικό-κοινωνικών παραγόντων που καθορίζουν το επίπεδο υγείας.

Στην Ελλάδα, σε αντίθεση με άλλες χώρες της ΕΕ, δεν υπάρχουν επιδημιολογικές μελέτες αντιπροσωπευτικές του πληθυσμού που να μας δίνουν έγκυρα στοιχεία για την κατάσταση της υγείας του πληθυσμού. Το κενό αυτό φιλοδοξεί να καλύψει η E.ΜΕ.ΝΟ.

Η E.ΜΕ.ΝΟ συνιστά μία μελέτη με ιδιαίτερη σημασία για τη χώρα, καθώς πραγματοποιείται με τη συνεργασία των πολιτών, στοχεύει στην καλύτερη υγεία όλων και αισιοδοξεί να συμβάλλει καθοριστικά στο σχεδιασμό μίας εθνικής στρατηγικής για την υγεία.

1.2 Το δείγμα

Η επιλογή του δείγματος είναι βασισμένη σε πολυσταδιακή στρωματοποιημένη κατά συστάδες δειγματοληψία των κατοίκων της Ελλάδας, όπως έχουν καταγραφεί από την Εθνική Στατιστική Υπηρεσία της Ελλάδος (ΕΣΥΕ) κατά την απογραφή του 2011. Πιο συγκεκριμένα η διαδικασία επιλογής περιλαμβάνει τρία στάδια. Στο πρώτο στάδιο ο Ελληνικός πληθυσμός διαιρείται σε 45 στρώματα ανά γεωγραφική περιοχή και βαθμό αστικοποίησης. Η γεωγραφική περιοχή αντανακλά διαφορές σε ποικίλους παράγοντες συμπεριλαμβανομένων των διατροφικών συνηθειών (ανάλογα με τις ανά περιοχή πηγές εισοδήματος και τα παραγόμενα προϊόντα) του κλίματος και τις διαθέσιμες υπηρεσίες υγείας. Ο διαχωρισμός της Ελλάδας σε περιοχές είναι ο εξής: Ανατολική Μακεδονία και Θράκη, Κεντρική Μακεδονία, Θεσσαλονίκη, Δυτική Μακεδονία, Ήπειρος, Θεσσαλία, Δυτική Ελλάδα, Κεντρική Ελλάδα, Ιόνια Νησιά, Βόρειο Αιγαίο, Νότιο Αιγαίο, Κρήτη, Πελοπόννησος, Αθήνα και Πειραιάς, και υπόλοιπη Αττική. Ο βαθμός αστικοποίησης αντανακλά τις διαφορές στον τρόπο και στην ποιότητα ζωής, στην κοινωνικοοικονομική κατάσταση και στην πληροφόρηση στα θέματα υγείας. Ο διαχωρισμός με βάση τον βαθμό αστικοποίησης είναι ο εξής: αστικές, ήμι-αστικές και αγροτικές περιοχές.

Στο δεύτερο στάδιο κάθε στρώμα χωρίζεται ανά οικοδομικό τετράγωνο (συστάδες). Η πιθανότητα επιλογής των συστάδων είναι ανάλογη του μεγέθους του στρώματος στο οποίο ανήκουν και στο μέγεθος των ίδιων των συστάδων. Στο τρίτο στάδιο γίνεται η επιλογή νοικοκυριών στα επιλεγμένα οικοδομικά τετράγωνα μέσω συστηματικής δειγματοληψίας και μέσα σε κάθε νοικοκυριό τυχαία επιλογή ενός ατόμου (Cochran 1997, Scheaffer RL 1996).Το δειγματοληπτικό σχήμα διασφαλίζει την τυχαιότητα και την αντιπροσωπευτικότητα του δείγματος στα βασικά χαρακτηριστικά του πληθυσμού. Το προτεινόμενο ελάχιστο μέγεθος δείγματος ανά χώρα είναι 4000 άτομα (Tolonen et al. 2008). Με βάση την αναμενόμενη χρηματοδότηση, ένα δείγμα μεγέθους 5000 ατόμων είναι λογικό. Στην μελέτη ΑΤΤΙΚΑ (πληθυσμιακή μελέτη που κάλυπτε την Αττική) ο επιπολασμός της υπερχοληστερολαιμίας, της υπέρτασης και του σακχαρώδη διαβήτη εκτιμήθηκε στο 40%, 32% και 7% αντιστοίχως (Panagiotakos et al. 2002) . Με 5000 συμμετέχοντες τέτοιος επιπολασμός μπορεί να εκτιμηθεί με τυχαία σφάλματα 0.98%, 0.79% και 0.38% αντίστοιχα για όλη την χώρα με βάση την απογραφή του 2001. Προκειμένου να μπορούν να εκτιμηθούν και λιγότερο συχνοί παράγοντες κινδύνου και ταυτόχρονα η μελέτη να είναι εφικτή από οικονομικής πλευράς το δείγμα επιλέχθηκε να είναι της τάξεως των 6000 ατόμων.

1.3 Τα δεδομένα

Τα δεδομένα συλλέγονται από τις παρακάτω πηγές:

1. Συνέντευξη:
   Εκπαιδευμένοι συνεντευκτές με χρήση δομημένων ερωτηματολογίων, συλλέγουν δεδομένα σχετικά με δημογραφικούς και κοινωνικοοικονομικούς παράγοντες, εθισμούς (καπνός, αλκοόλ, ναρκωτικά), ιατρικό ιστορικό, έκθεση σε περιβαλλοντικούς ή εργασιακούς κινδύνους. Δεδομένου ότι έμφαση δίνεται σε καρδιαγγειακά και αναπνευστικά νοσήματα, το ιατρικό ιστορικό και η έκθεση σε περιβαλλοντικούς ή εργασιακούς κινδύνους έχουν δομηθεί καταλλήλως ώστε να εξασφαλιστεί ότι θα συλλεχθούν όλα τα απαραίτητα δεδομένα για την αξιολόγηση του βραχυπρόθεσμου και μεσοπρόθεσμου κινδύνου ανάπτυξης/εξέλιξης καρδιαγγειακών και αναπνευστικών νοσημάτων. Συγκεκριμένα, αναπτύχθηκαν σχετικά σύντομα ερωτηματολόγια που αξιολογούν:
   α) διατροφή και τον βαθμό προσήλωσης στην μεσογειακή διατροφή,
   β) φυσική δραστηριότητα,
   γ) γενική κατάσταση της υγείας και χρήση υπηρεσιών υγείας,
   δ) φαρμακευτικές αγωγές και προληπτικά μέτρα.

2. Εξέταση ανθρωπομετρικών και σωματομετρικών χαρακτηριστικών:
   Με χρήση τυποποιημένων οργάνων και ειδικά διαμορφωμένων τυποποιημένων διαδικασιών, πραγματοποιούνται σωματικές μετρήσεις από εκπαιδευμένο ιατρικό προσωπικό. Τα δεδομένα που συλλέγονται περιλαμβάνουν το σωματικό βάρος, ύψος, περιφέρεια μέσης, πίεση και παλμό.

3. Σπιρομέτρηση:
   Χρησιμοποιούνται τυποποιημένα σπιρόμετρα και ειδικά διαμορφωμένες διαδικασίες. Για την επιδημιολογική διερεύνηση της Χρόνιας Αποφρακτικής Πνευμονοπάθειας (ΧΑΠ) χρησιμοποιούνται τα διαχωριστικά σημεία της σπιρομέτρησης για την διάγνωση και την σοβαρότητα όπως αυτά ορίζονται από την Global Initiative for Chronic Obstructive Lung Disease (GOLD) (Rabe KF 2007). Ενώ τα αντίστοιχα από τις οδηγίες των European Respiratory Society (ERS) /American Thoracic Society (ATS) , χρησιμοποιούνται για την διάγνωση και την αντιμετώπιση των ασθενών με ΧΑΠ (Celli BR 2004). Η σπιρομέτρηση πραγματοποιείται σύμφωνα με τις οδηγίες των ATS/ERS .

4. Αιματολογικές εξετάσεις:
   Τα επίπεδα γλυκόζης και χοληστερόλης στο αίμα μετρούνται με χρήση τυποποιημένων διαδικασιών. Με την γραπτή συναίνεση των συμμετεχόντων, θα δημιουργηθεί μια τράπεζα αίματος για μελλοντική χρήση, με την χρήση των δειγμάτων για τεστ DNA να απαγορεύεται ρητώς, εκτός εάν υπάρξει ειδική σχετική συναίνεση.

5. Ατμοσφαιρική ρύπανση:
   Στις περισσότερες αστικές περιοχές έχουν εγκατασταθεί κέντρα παρακολούθησης της ατμοσφαιρικής ρύπανσης. Τέτοια κέντρα έχουν επίσης εγκατασταθεί κοντά σε σημεία-πηγές σε κάποιες μη αστικές περιοχές. Τέτοια δεδομένα μαζί με τις αντίστοιχες μέσες ημερήσιες θερμοκρασίες θα ζητηθούν από τις σχετικές αρχές. Επιπλέον λόγω του γεγονότος ότι οι συγκεντρώσεις Όζοντος είναι ένα πρόβλημα στην Ελλάδα λόγω τής ηλιοφάνειας, και ότι το Διοξείδιο του Αζώτου είναι ένας καλός δείκτης κυκλοφοριακής ρύπανσης, δεδομένα και για αυτούς τους δύο δείκτες θα ζητηθούν.

1.4 Σκοπός της μελέτης

Η ΕΜΕΝΟ είναι μια συγχρονική επιδημιολογική μελέτη εξέτασης υγείας (Health Examination Study - HES) με κύριους σκοπούς:

1. Να περιγράψει το φάσμα νοσηρότητας στον Ελληνικό ενήλικο πληθυσμό.

2. Να εκτιμήσει:
   α) Τον επιπολασμό των βασικών κινδύνων υγείας για χρόνιες παθήσεις και να ελέγξει για πιθανές διαφορές ανά γεωγραφική περιοχή, ηλικία, κοινωνικό-οικονομικούς παράγοντες και άλλα χαρακτηριστικά του υπό μελέτη πληθυσμού.
   β)Τον βαθμό προσήλωσης στην μεσογειακή διατροφή και τα επίπεδα φυσικής δραστηριότητας, και να αξιολογήσει συγχρονικά την επίδρασή τους στον επιπολασμό των νοσημάτων.
   γ) Τον επιπολασμό των χρόνιων αναπνευστικών νοσημάτων, ειδικά της ΧΑΠ και του άσθματος.
   δ) Τον κίνδυνο των καρδιαγγειακών νοσημάτων και πιθανούς παράγοντες κινδύνου.

3. Να αξιολογήσει:
   α) Την γενική κατάσταση υγείας του πληθυσμού και τον συσχετισμό της με δείκτες υγείας.
   β) Την χρήση υπηρεσιών υγείας, φαρμακευτικών αγωγών και μέτρων πρόληψης.
   γ) Την σχέση μεταξύ της έκθεσης στην ατμοσφαιρική ρύπανση και την πνευμονική λειτουργία.

4. Να εκτιμήσει μελλοντική επιβάρυνση της υγείας και να αξιολογήσει τις μελλοντικές ανάγκες του συστήματος υγείας.

1.5 Η Hprolipsis

Ταυτόχρονα με την ΕΜΕΝΟ διεξάγεται και η μελέτη Hprolipsis. Η Hprolipsis είναι η πρώτη Εθνική Επιδημιολογική Μελέτη των λοιμωδών νοσημάτων ηπατίτιδα Β (HBV), ηπατίτιδα C (HCV) και της HIV λοίμωξης. Στοχεύει:
α) στη συλλογή έγκυρων δεδομένων για τη συχνότητα, τους παράγοντες κινδύνου και τις στάσεις και γνώσεις για τις ηπατίτιδες Β και C και τον HIV στο γενικό ενήλικο πληθυσμό καθώς και σε ευάλωτους πληθυσμούς (Τσιγγάνοι/Ρομά και Μετανάστες),
β) στην υλοποίηση δράσεων ενημέρωσης και ευαισθητοποίησης των πληθυσμών στόχων,
γ) στον εμβολιασμό των επίνοσων ατόμων των ευάλωτων πληθυσμών και
δ) στην ατομική συμβουλευτική ορολογικών ατόμων και στην διασύνδεσή τους με το δημόσιο σύστημα παροχής υγείας.

Κεφάλαιο 2

Μη ανταπόκριση σε επίπεδο Ατόμου

2.1 Σφάλματα στις πληθυσμιακές μελέτες

Στις πληθυσμιακές μελέτες τα δεδομένα συλλέγονται από ένα μικρό υποσύνολο του πληθυσμού. Το δείγμα παρέχει πληροφορίες που αφορούν μόνο το δείγμα, όμως, είναι δυνατό να εξαχθούν συμπεράσματα τα οποία αφορούν στον πληθυσμό. Με κατάλληλο δειγματοληπτικό σχεδιασμό, και υπό την προϋπόθεση ότι δεν υπάρχουν άλλα προβλήματα, είναι δυνατόν να υπολογιστούν αξιόπιστες εκτιμήσεις χαρακτηριστικών του πληθυσμού.

Οι εκτιμήσεις αυτές θα διαφέρουν από δείγμα σε δείγμα, ποτέ δεν θα είναι οι ακριβείς τιμές του πληθυσμού. Πάντα θα υπάρχει σφάλμα, το οποίο μπορεί να οφείλεται σε πολλούς λόγους. Δύο μεγάλες κατηγορίες μπορούν να διακριθούν. Δειγματοληπτικά και μη δειγματοληπτικά σφάλματα (Jelke Bethlehem 2004):

- Τα δειγματοληπτικά σφάλματα εισάγονται εξαιτίας του σχεδιασμού. Οφείλονται στο γεγονός ότι οι εκτιμήσεις βασίζονται σε δείγμα και όχι στο σύνολο του πληθυσμού. Το δείγμα επιλέγεται με κάποια τυχαία διαδικασία. Κάθε φορά που εφαρμόζεται μια τέτοια διαδικασία το δείγμα θα αποτελείται από διαφορετικά άτομα. Έτσι, σε επαναλήψεις μιας μελέτης, οι εκτιμήσεις θα διαφέρουν. Η επίδραση τέτοιου είδους σφάλματος μπορεί να ελεγχθεί μέσω του σχεδιασμού. Για παράδειγμα με αύξηση του μεγέθους δείγματος ή με πιθανότητες επιλογής ανάλογες σε κάποια καλά ορισμένη βοηθητική μεταβλητή.

- Μη-δειγματοληπτικά σφάλματα μπορεί να προκύψουν ακόμα και εάν όλος ο πληθυσμός μελετάται. Αυτά προκύπτουν κατά τη συλλογή των δεδομένων. Σημαντική πηγή μη δειγματοληπτικού σφάλματος αποτελεί η μη ανταπόκριση. Υπάρχουν διάφοροι λόγοι για τους οποίους μπορεί να προκύψει μη ανταπόκριση όπως άρνηση συνεργασίας, απουσία από το σπίτι κατα την ώρα της επίσκεψης, πρόβλημα γλώσσας, ασθένεια κ.α.

Η εγκυρότητα της στατιστικής συμπερασματολογίας μιας πληθυσμιακής μελέτης, εξαρτάται από τον βαθμό στον οποίο το δείγμα είναι αντιπροσωπευτικό του πληθυσμού από τον οποίο προέρχεται. Αυξανομένων των ποσοστών μη ανταπόκρισης είναι εύλογο να αναρωτηθεί κανείς εάν εκείνοι οι οποίοι ανταποκρίθηκαν τελικά είναι αντιπροσωπευτικοί του γενικού πληθυσμού. Εάν η μη ανταπόκριση συνδέεται με τις μεταβλητές ενδιαφέροντος τότε μάλλον το δείγμα μας παρέχει μια διαστρεβλωμένη εικόνα των χαρακτηριστικών του πληθυσμού. Έχει ενδιαφέρον να εξετάσουμε το πως η μη ανταπόκριση μπορεί να προκαλέσει μεροληψία στις εκτιμήσεις μας (μέσους, ποσοστά κ.α),

από διάφορες σκοπιές. Στην πραγματικότητα το πρόβλημα είναι πρόβλημα ελλείπουσων τιμών. Οι κύριοι μηχανισμοί οι οποίοι δημιουργούν ελλείπουσες τιμές σύμφωνα με τον Rubin είναι οι εξής (Rubin 1987):

Κύριοι Μηχανισμοί Παραγωγής ελλειπουσών τιμών

- Εντελώς τυχαίος (Missing Completely at Random , MCAR): όταν η πιθανότητα να λείπουν δεδομένα είναι ανεξάρτητη των παρατηρηθέντων αλλά και των μη παρατηρηθέντων τιμών. Σε αυτή την περίπτωση οι τιμές που λείπουν αποτελούν τυχαίο δείγμα των δεδομένων και η ανάλυση που δεν τις λαμβάνει υπόψιν δίνει αμερόληπτες εκτιμήσεις. Στην πράξη αυτό σπάνια προκύπτει.
- Τυχαίος (Missing at Random , MAR): όταν η πιθανότητα να λείπουν δεδομένα δεν εξαρτάται από μη παρατηρηθέντα δεδομένα αλλά μπορεί να εξαρτάται από τα παρατηρηθέντα. Σε αυτή την περίπτωση τα δεδομένα που λείπουν δεν περιέχουν επιπλέον πληροφορία εάν λάβουμε υπόψιν μας τα παρατηρηθέντα δεδομένα. Η στατιστική ανάλυση που περιορίζεται στα άτομα τα οποία δεν έχουν ελλείπουσες τιμές (complete cases), θα δώσει μεροληπτικά αποτελέσματα.
- Όχι τυχαίος (Missing Not at Random , MNAR): όταν η πιθανότητα να λείπουν δεδομένα εξαρτάται από μη παρατηρηθέντα δεδομένα.

Συνήθως υποθέτουμε ότι τα δεδομένα είναι MAR αν και αυτή η υπόθεση δεν είναι δυνατό να ελεγχθεί.

2.2 Σφάλμα μη ανταπόκρισης

Μια πρώτη προσέγγιση υποθέτει ότι η ανταπόκριση ή μη είναι ένα από τα χαρακτηριστικά του ατόμου, χωρίζοντας έτσι τον πληθυσμό σε εκείνους οι οποίοι πάντα θα ανταποκρίνονται (responders)και σε εκείνους που πάντα δεν θα ανταποκρίνονται (non-repsonders). Με βάση τα παραπάνω ορίζουμε τις ποσότητες: $\mathbf{N}$ =ο συνολικός πληθυσμός
N_1= Ο συνολικός αριθμός των ατόμων που ανταποκρίνονται
N_2= Ο συνολικός αριθμός των ατόμων που δεν ανταποκρίνονται
$\bar{Y_1}$ = ο μέσος ενός χαρακτηριστικού $\mathbf{Y}$ στον N_1 πληθυσμό
$\bar{Y_2}$ = ο μέσος ενός χαρακτηριστικού $\mathbf{Y}$ στον N_2 πληθυσμό

Συνεπώς ο μέσος ενός χαρακτηριστικού $\mathbf{Y}$ στον συνολικό πληθυσμό, προκύπτει ως ένας σταθμισμένος μέσος των $\bar{Y_1}$ και $\bar{Y_2}$ ως εξής:

$$\bar{Y} = \frac{N_1\bar{Y_1} + N_2\bar{Y_2}}{\mathbf{N}}$$

Εάν δεν συνυπολογίσουμε εκείνους που δεν ανταποκρίθηκαν τότε θα έχουμε:

$$E(Y) = \bar{Y_1}$$

συνεπώς μπορούμε να υπολογίσουμε το σφάλμα του εκτιμητή ως εξής:

$$B(\bar{Y}) = \bar{Y_1} - \bar{Y} = (\frac{N_2}{\mathbf{N}})(\bar{Y_1} - \bar{Y_2}) \tag{2.2.1}$$

Εξετάζοντας την παραπάνω σχέση το συμπέρασμα είναι ότι το σφάλμα θα είναι σχετικά μικρό εάν η αναλογία εκείνων που δεν ανταποκρίνονται στον συνολικό πληθυσμό είναι μικρή, ή εάν η διαφορά στις μέσες τιμές μεταξύ εκείνων που ανταποκρίνονται και εκείνων που όχι είναι μικρή (Paul S.Levy 2008).

Τέτοιου είδους σφάλμα μπορούμε να έχουμε και για άλλες παραμέτρους πέραν της μέσης τιμής. Για παράδειγμα, όταν ενδιαφερόμαστε να εκτιμήσουμε τον συνολικό αριθμό των ατόμων του πληθυσμού οι οποίοι έχουν ένα χαρακτηριστικό (πληθυσμιακό ολικό, population total), παρουσία μη ανταπόκρισης, οι εκτιμήσεις θα είναι υποεκτιμημένες. Η μη ανταπόκριση επίσης μπορεί να επηρεάσει και τους συντελεστές της παλινδρόμησης αν ενδιαφερόμαστε να διερευνήσουμε πως κάποιες μεταβλητές παράγοντες επηρεάζουν την μεταβλητή απόκρισης. Εάν η σχέση μεταξύ των ανεξάρτητων μεταβλητών X και της μεταβλητής απόκρισης Υ είναι πιο ισχυρή ανάμεσα στους ανταποκριθέντες από ότι ανάμεσα στους μη ανταποκριθέντες, τότε ο συντελεστής της παλινδρόμησης θα υπερεκτιμηθεί (Groves 1998).

Μια μελέτη προσομοίωσης (Jones 1995) έδειξε ότι εάν δεν υπάρχει καμία διαφορά μεταξύ ανταποκριθέντων και μη, χαμηλό ποσοστό ανταπόκρισης δεν αποτελεί πρόβλημα. Με μικρές διαφορές όμως, ένα ποσοστό ανταπόκρισης 50% δίνει στον ερευνητή μόλις 48% πιθανότητα να εκτιμήσει σωστά τον πληθυσμιακό μέσο, ακόμα και μέσα στο ευρύ 95% διάστημα εμπιστοσύνης. Για μεγάλες διαφορές, ένα ποσοστό ανταπόκρισης της τάξης του 90% θα ήταν απαραίτητο ώστε να διασφαλιστεί η σωστή εκτίμηση του πληθυσμιακού μέσου (Jones 1996, Gary 2007).

Από μια άλλη, πιο στοχαστική σκοπιά, η προσέγγιση σύμφωνα με την οποία η ανταπόκριση είναι αναπόσπαστη ιδιότητα του ατόμου, μπορεί να θεωρηθεί λανθασμένη. Αντιθέτως, θα ήταν πιο λογικό να αναθέσουμε στο κάθε άτομο μια πιθανότητα να ανταποκριθεί ή όχι ανάλογα με τις περιστάσεις (Lessler & Kalsbeek 1992). Αναπαριστώντας αυτή την πιθανότητα (propensity) με ρ_i προκύπτει μια άλλη έκφραση του σφάλματος του μέσου (Bethlehem 2002):

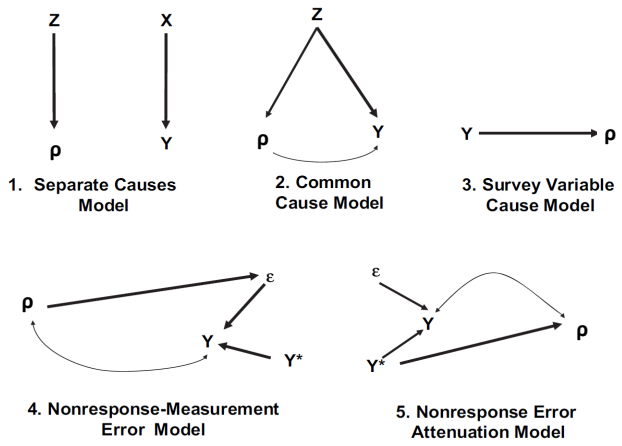
$$B(\bar{Y}) \approx \frac{\sigma_{Y\rho}}{\bar{\rho}} \tag{2.2.2}$$

όπου:
$\sigma_{Y\rho}$ = η συνδυακύμανση μεταξύ της υπό μελέτης μεταβλητής Y και της πιθανότητας ανταπόκρισης ρ
και
$\bar{\rho}$= η μέση πιθανότητα ανταπόκρισης στον πληθυσμό (επαναλαμβανόμενες δειγματοληψίες με τον ίδιο σχεδιασμό).

Σύμφωνα με αυτή την έκφραση, το σφάλμα προκύπτει ως συνάρτηση της συσχέτισης της πιθανότητας ανταπόκρισης με την μεταβλητή ενδιαφέροντος. Αξίζει να σημειωθεί ότι ακόμα και στην ίδια μελέτη, διαφορετικές εκτιμήσεις μπορεί να υπόκεινται σε διαφορετικό σφάλμα μη ανταπόκρισης. Εάν η συσχέτιση είναι μικρή, η εκτίμηση μας μπορεί να είναι ελεύθερη σφάλματος λόγω μη ανταπόκρισης. Στην αντίθετη περίπτωση, εάν η συσχέτιση είναι μεγάλη το σφάλμα θα είναι ανάλογα μεγάλο.

Βλέποντας το θέμα από μια αιτιολογική σκοπιά, μπορούμε να εξερευνήσουμε κάποιους από τους μηχανισμούς που είναι υπεύθυνοι για την δημιουργία τέτοιων συσχετίσεων και τέτοιων σφαλμάτων. Θα δούμε πέντε αιτιολογικά μοντέλα (Σχήμα 2.1).

Το πρώτο μοντέλο, ξεχωριστών αιτιών (Separate Causes Model), περιγράφει μια κατάσταση στην οποία δεν δημιουργείται καμία συσχέτιση μεταξύ της μεταβλητής Y και της πιθανότητας ανταπόκρισης ρ. Σε αυτή την περίπτωση υπάρχει ένα σετ αιτιών Z της πιθανότητας ανταπόκρισης πλήρως διαχωρισμένων και ανεξάρτητων των μεταβλητών που επηρεάζουν την μεταβλητή ενδιαφέροντος Y . Δεν υπάρχει εδώ σφάλμα οφειλόμενο στην μη ανταπόκριση ανεξάρτητα από το μέγεθος της. Αυτό το σενάριο είναι ιδιαίτερα απλουστευμένο και σπάνιο να εμφανιστεί στην πράξη και αποτελεί παράδειγμα του MCAR μηχανισμού παραγωγής ελλείπουσων τιμών.



Σχήμα 2.1: Πέντε αιτιολογικά μοντέλα της σχέσης της πιθανότητας ανταπόκρισης ρ, της μεταβλητής απόκρισης Υ, της πραγματικής τιμής Υ^* και άλλων μεταβλητών (X,Z), και του σφάλματος μέτρησης ε, με διαφορετικές συνέπειες στο σφάλμα μη ανταπόκρισης (Groves 2006).

Το δεύτερο μοντέλο, κοινής αιτίας (Common Cause Model), δημιουργεί συσχέτιση μεταξύ των δύο χαρακτηριστικών, της μεταβλητής Y και της πιθανότητας ανταπόκρισης ρ, λόγω της κοινής τους αιτίας Z (ή ένα σύνολο μεταβλητών Z). Σε αυτό το σενάριο, εάν οι μεταβλητές Z μετρηθούν, μπορεί η επίδρασή τους να εξαλειφθεί με ανάλογες στατιστικές διορθώσεις. Αυτό το μοντέλο αντιστοιχεί στον MAR μηχανισμό.

Στο τρίτο μοντέλο, η υπό μελέτη μεταβλητή Y αποτελεί την αιτία της πιθανότητας μη ανταπόκρισης (Survey Variable Cause Model) δημιουργώντας έτσι μια απευθείας συσχέτιση. Αυτή η περίπτωση χαρακτηρίζεται ως πληροφοριακή μη ανταπόκριση και αποτελεί παράδειγμα του MNAR μηχανισμού παραγωγής ελλείπουσων τιμών.

Το τέταρτο μοντέλο, Nonresponse-Measurement Error Model , περιγράφει ένα συνδυασμό μη ανταπόκρισης και σφάλματος μέτρησης. Η πιθανότητα ανταπόκρισης ρ επιδρά στο μέγεθος του σφάλματος μέτρησης ε, το οποίο σχετίζεται με την μεταβλητή Y. Ισχύει δηλαδή $Y_i = Y_i^* + \varepsilon_i$, το οποίο σημαίνει ότι η παρατηρούμενη τιμή Y_i ισούται με την πραγματική τιμή Y_i^* συν έναν όρο σφάλματος ε_i. Η συσχέτιση μεταξύ Y και ρ δημιουργείται μέσω της αιτιολογικής σχέσης μεταξύ ρ και ε. Και σε αυτή την περίπτωση εάν γνωρίζουμε τον μηχανισμό με τον οποίο δημιουργούνται τα σφάλματα μέτρησης μπορούμε να εξαλείψουμε το σφάλμα μη ανταπόκρισης.

Το πέμπτο μοντέλο,Nonresponse Error Attenuation Model, περιγράφει μια περίπτωση κατά την οποία δεν δημιουργείται σφάλμα μη ανταπόκρισης. Ισχύει ότι $Y_i = Y_i^* + \varepsilon_i$ όπου Y_i η παρατηρούμενη τιμή, Y_i^* η πραγματική τιμή, με την υπόθεση ότι δεν υπάρχει συνδυακύμανση μεταξύ της πραγματικής τιμής και του όρου σφάλματος. Σε αυτήν την περίπτωση η οποιαδήποτε συνδυακύμανση μεταξύ της πραγματικής τιμής Y_i^* και της πιθανότητας ανταπόκρισης ρ εξομαλύνεται από την διακύμανση του σφάλματος μέτρησης.

Συνοψίζοντας, το μέγεθος του σφάλματος λόγω μη ανταπόκρισης, συνδέεται άμεσα με την συσχέτιση της μεταβλητής ενδιαφέροντος με την πιθανότητα ανταπόκρισης. Ανάμεσα στις διάφορες μεταβλητές της ίδιας έρευνας η επίδραση αυτή, κατά πάσα πιθανότητα, θα ποικίλει. Δεν υπάρχει απλή σχέση μεταξύ του ποσοστού μη ανταπόκρισης και του σφάλματος λόγω αυτής. Ιδανικά θα θέλαμε να πετύχουμε το υψηλότερο δυνατό ποσοστό ανταπόκρισης, γνωρίζοντας όμως ότι παρά τις προσπάθειες κάποιο ποσοστό μη ανταπόκρισης θα παραμείνει, οι ερευνητές θα πρέπει να αναζητούν

άλλους τρόπους μείωσης της επίδρασης της μη ανταπόκρισης.

2.3 Διόρθωση του σφάλματος Μη Ανταπόκρισης

Το πρώτο είδος μη-ανταπόκρισης που καλούμαστε να αντιμετωπίσουμε είναι η μη-ανταπόκριση σε επίπεδο ατόμου (unit nonresponse). Αυτή ορίζεται ως η αποτυχία απόκτησης ανταπόκρισης-πληροφορίας από ένα επιλεγμένο στο δείγμα άτομο (Brick 2013). Αυτή είναι μια μορφή ελλειπουσών τιμών στα δεδομένα. Στις πληθυσμιακές μελέτες η ανάλυση γίνεται με μεθόδους βασισμένες στον σχεδιασμό. Αυτές οι μέθοδοι όμως απαιτούν πλήρη δεδομένα.

Σημαντικό ρόλο παίζει η διαθεσιμότητα κάποιας, έστω μερικής, πληροφορίας για τους μη ανταποκριθέντες. Δύο τρόποι υπάρχουν για την απόκτηση αυτής της πληροφορίας (Jelke Bethlehem 2004):

- Προσπάθεια να ληθφεί πληροφορία από τους μη ανταποκριθέντες. Αυτό απαιτεί επιπλέον προσπάθεια επικοινωνίας μαζί τους.
- Προσπάθεια να ληθφεί πληροφορία για τους μη ανταποκριθέντες. Αυτό απαιτεί λήψη πληροφορίας από εξωτερικές πηγές.

Η πρώτη αντιμετώπιση περιγράφεται παρακάτω. Η δεύτερη ανήκει σε μια ευρεία κατηγορία αντιμετώπισης ελλειπουσών τιμών, με πολλές διαφορετικές μεθόδους. Αυτές οι μέθοδοι βασίζονται στην στάθμιση και είναι εφαρμόσιμες τόσο στην μη ανταπόκριση ατόμου όσο και στην μη ανταπόκριση ερώτησης. Λόγω του ότι αποτελεί σημαντική κατηγορία θα περιγραφεί ξεχωριστά.

2.3.1 Επιλογή δείγματος από μη ανταποκριθέντες.

Η πρώτη μέθοδος που θα περιγράψουμε αντιμετωπίζει τους μη αντοποκριθέντες σαν ένα στρώμα του πληθυσμού (Hansen & Hurwitz 1946). Στην πρώτη φάση της δειγματοληψίας ένα ποσοστό των επιλεγμένων ατόμων ανταποκρίνονται ενώ ένα άλλο όχι. Ο ερευνητής έχει δύο επιλογές. Είτε να προσπαθήσει να επικοινωνήσει ξανά με όλους όσους δεν ανταποκρίθηκαν είτε να προσπαθήσει να επικοινωνήσει ξανά με ένα τυχαίο δείγμα όσων δεν ανταποκρίθηκαν. Η μέθοδος που περιγράφεται εδώ (Paul S.Levy 2008) αφορά στην δεύτερη περίπτωση. Ακόμα και αν όχι όλοι όσοι επιλεγούν στο τυχαίο αυτό δείγμα, επιλέξουν να ανταποκριθούν σε αυτή τη δεύτερη φάση, ο ερευνητής θα έχει καταφέρει να συλλέξει πολύτιμες πληροφορίες για αυτό το στρώμα του πληθυσμού. Στην συνέχεια η ανάλυση γίνεται με τις κλασσικές τεχνικές για στρωματοποιημένη δειγματοληψία.
Συγκεκριμένα αν ορίσουμε ως:
N =το μέγεθος του πληθυσμού
n = το μέγεθος του δείγματος στην πρώτη φάση
n_1 = ο αριθμός των ανταποκριθέντων στην πρώτη φάση
$n_2 = n - n_1$ = ο αριθμός των μη ανταποκριθέντων στην πρώτη φάση
n_2^* = το μέγεθος του τυχαία επιλεγμένου δείγματος στη δεύτερη φάση
n_2' = ο αριθμός των ανταποκριθέντων στη δεύτερη φάση

$\bar{Y_1} = \frac{\sum_{i=1}^{n_1} Y_i}{n_1}$ = η μέση τιμή του χαρακτηριστικού Y ανάμεσα στους n_1 ανταποκριθέντες της πρώτης φάσης

$\bar{Y_2} = \frac{\sum_{i=1}^{n_2'} Y_i}{n_2'}$ = η μέση τιμή του χαρακτηριστικού Y ανάμεσα στους n_2' ανταποκριθέντες της δεύτερης φάσης.

Ένας εκτιμητής της άγνωστης μέσης τιμής του πληθυσμού δίνεται παρακάτω:

$$\bar{Y} = \frac{n_1 \bar{Y_1} + n_2 \bar{Y_2}}{n} \tag{2.3.1}$$

Όσο ο αριθμός των ανταποκριθέντων στην δεύτερη φάση n_2' πλησιάζει τον αριθμό n_2^* των ατόμων που επιλέχθηκαν στην δεύτερη φάση, ο εκτιμητής $\bar{Y}$ γίνεται αμερόληπτος.

Σε μια μελέτη προσομοίωσης οι Jenkins et al το 2007 (Jenkins et al. 2008) έδειξαν ότι για αλλαγή του ποσοστού ανταπόκρισης από 30% σε 60% οι εκτιμήσεις του επιπολασμού για τέσσερα χρόνια νοσήματα ταυτίζονταν με τις παραμέτρους του πληθυσμού, αν εφαρμοζόταν η μέθοδος των δύο σταδίων. Οι αντίστοιχες εκτιμήσεις που προήλθαν από ανάλυση με μεθόδους στάθμισης, διέφεραν από εκείνες του πληθυσμού όχι όμως περισσότερο από 0.2%. Οι ίδιοι συγγραφείς αναφέρουν, όμως, ότι οι διακυμάνσεις της μεθόδου των δύο σταδίων ήταν σημαντικά μεγαλύτερες από εκείνες που αποκτήθηκαν από τις μεθόδους στάθμισης. Αυτό αναφέρεται και ως το κύριο μειονέκτημα αυτής της μεθοδολογίας. Επιπλέον για την εφαρμογή αυτής της μεθόδου θα πρέπει να συνυπολογιστεί το κόστος σε χρήμα και σε χρόνο.

2.4 Στάθμιση Αντίστροφης Πιθανότητας

Οι μέθοδοι στάθμισης μπορούν να χρησιμοποιηθούν για την διόρθωση μη ανταπόκρισης τόσο σε επίπεδο ατόμου όσο και σε επίπεδο ερώτησης. Είθισται όμως, να χρησιμοποιούνται κυρίως στην περίπτωση της μη ανταπόκρισης ατόμου ενώ για την μη ανταπόκριση ερώτησης συνηθέστερες είναι οι μέθοδοι πολλαπλής αντικατάστασης δεδομένων (multiple imputations, MI.) Σε αυτή την ενότητα γίνεται περιγραφή τέτοιων μεθόδων ενώ σε παρακάτω κεφάλαιο θα γίνει περιγραφή και της MI.

2.4.1 Χρήση Βαρών στις Πληθυσμιακές έρευνες

Συχνά, τα άτομα ενός δείγματος μιας πληθυσμιακής μελέτης συνοδεύονται από κάποια βάρη με σκοπό οι σταθμισμένες παρατηρήσεις να είναι όσο το δυνατό πιό αντιπροσωπευτικές του πληθυσμού αναφοράς. Η δημιουργία των βαρών αυτών γίνεται σε μια σειρά βημάτων και χρησιμοποιείται για διόρθωση άνισης πιθανότητας επιλογής, μη ανταπόκρισης, μη κάλυψης καθώς και για αποκλίσεις από γνωστές τιμές του πληθυσμού (Brick & Kalton 1996, Kalton et al. 1998).

Αρχικά για την διόρθωση για άνιση πιθανότητα επιλογής η κάθε παρατήρηση σταθμίζεται, συνήθως, με την αντίστροφη πιθανότητα επιλογής στο δείγμα ή με κάποια ποσότητα ανάλογη με αυτήν. Η άνιση επιλογή είναι συνήθως θέμα σχεδιασμού άρα γνωρίζουμε την πιθανότητα επιλογής. Συνεπώς η δημιουργία αυτών των βαρών είναι σχετικά απλή.

Στην συνέχεια για να γίνει διόρθωση για μη ανταπόκριση ατόμου, τα αρχικά βάρη τροποποιούνται. Η βασική αρχή, είναι η αναγνώριση ατόμων που ανταποκρίθηκαν παρόμοια με εκείνα που δεν ανταποκρίθηκαν και η κατάταξη των ατόμων που ανταποκρίθηκαν σε ομάδες με παρόμοια χαρακτηριστικά με αυτά εκείνων που δεν ανταποκρίθηκαν. Αυτό γίνεται με χρήση βοηθητικών πληροφοριών που πιθανόν να διαθέτουν οι ερευνητές. Στη συνέχεια τα αρχικά βάρη αυξάνονται με τρόπο τέτοιο ώστε οι ανταποκριθέντες να είναι αντιπροσωπευτικοί των παρόμοιων τους μη ανταποκριθέντων. Στις περισσότερες περιπτώσεις η βοηθητική πληροφορία είναι περιορισμένη (συχνά γνωρίζουμε μόνο την συστάδα και το στρώμα). Σε αυτήν την περίπτωση τα άτομα του δείγματος χωρίζονται σε κλάσεις με βάση την πληροφορία αυτή. Αυτή η μέθοδος δουλεύει καλά όταν η βοηθητική πληροφορία είναι περιορισμένη. Σε αντίθετη περίπτωση είναι απαραίτητη η χρήση άλλων μεθόδων.

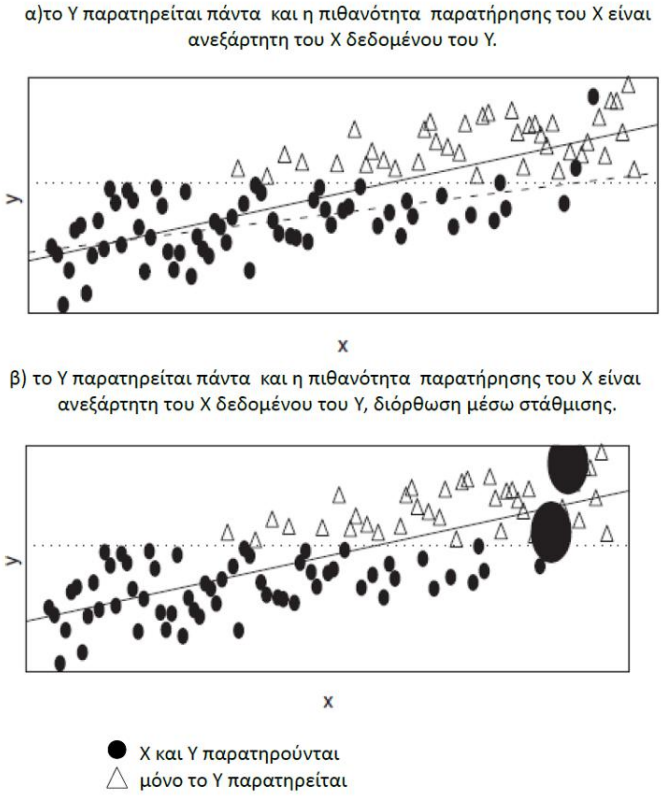
Επιπλέον μπορεί να γίνει χρήση γνωστών τιμών του πληθυσμού για περαιτέρω διορθώσεις, τροποποιώντας τα βάρη ώστε οι σταθμισμένες εκτιμήσεις να συμφωνούν με τιμές γνωστές για τον πληθυσμό, για κάποιες βασικές μεταβλητές. Αυτό το είδος διόρθωσης εξυπηρετεί δύο σκοπούς: αφενός διορθώνει για μη-κάλυψη και αφετέρου βελτιώνει την ακρίβεια των εκτιμήσεων. Μπορεί

όμως να χρησιμοποιηθεί και για διόρθωση για μη ανταπόκριση. Η διαδικασία αυτή αναφέρεται συχνά ως εκ των υστέρων στρωματοποίηση (poststratification). Όμως η κλασική εκ των υστέρων στρωματοποίηση υποθέτει πλήρη ανταπόκριση και τέλεια κάλυψη.

Σε τέτοια περίπτωση οι διορθώσεις αυτής της μεθόδου είναι σχετικά μικρές, και το κέρδος βρίσκεται στην αύξηση της ακρίβειας των εκτιμήσεων. Εάν όμως υπάρχει σημαντική μη ανταπόκριση και/ή μη κάλυψη, οι διορθώσεις μπορεί να είναι σημαντικές. Το σφάλμα μπορεί να μειωθεί όμως οι τυπικές αποκλίσεις μπορεί να αυξηθούν.

2.4.2 Διόρθωση μέσω στάθμισης

Μια επιλογή των ερευνητών όταν έχουν μη-πλήρη δεδομένα, είναι να κάνουν την ανάλυση με χρήση μόνο των πλήρων παρατηρήσεων (Complete Case analysis - CC analysis). Ας υποθέσουμε ότι έχουμε τα δεδομένα που φαίνονται στο σχήμα 2.2, αλλά μόνο οι παρατηρήσεις με κύκλο είναι πλήρεις. Σε όλες έχει παρατηρηθεί το Υ και η πιθανότητα παρατήρησης του X είναι ανεξάρτητη του X δεδομένου του Υ. Όταν το Υ είναι μικρότερο από κάποια τιμή το X παρατηρείται πάντα, ενώ όταν είναι πάνω από αυτή την τιμή το X σχεδόν πάντα λείπει. Αυτό αντιστοιχεί στην MAR υπόθεση. Η τιμή αυτή του Υ είναι ένα διαχωριστικό σημείο (διακεκομμένη οριζόντια γραμμή). Η παλινδρόμηση του Υ στο X, βασισμένη μόνο στις πλήρεις παρατηρήσεις δίνει λανθασμένη εκτίμηση της κλίσης (η συνεχόμενη γραμμή αναπαριστά την πραγματική κλίση ενώ η διακεκομμένη την εκτιμώμενη κλίση).



Σχήμα 2.2: Γραμμική παλινδρόμηση του Υ στο X. Η συνεχής γραμμή αναπαριστά την πραγματική κλίση ενώ η διακεκομμένη γραμμή την εκτιμώμενη κλίση (Seaman & White 2013).

Αν χρησιμοποιήσουμε μεθόδους στάθμισης η παλινδρόμηση θα γίνει πάλι μόνο στις πλήρεις παρατηρήσεις, αυτή τη φορά όμως, διαφορετικά βάρη θα δοθούν σε κάποιες παρατηρήσεις. Στο σχήμα 2.2, το δεύτερο διάγραμμα, δείχνει πως γίνεται η διόρθωση μέσω στάθμισης. Τα δεδομένα στα δύο διαγράμματα είναι ίδια. Οι δύο παρατηρήσεις που βρίσκονται πάνω από την τιμή του Υ που λειτουργεί ως διαχωριστικό σημείο, λαμβάνουν μεγάλα βάρη (μεγάλοι κύκλοι σημαίνουν μεγάλο βάρος). Με αυτό τον τρόπο οι δύο αυτές παρατηρήσεις επηρεάζουν κατά πολύ την εκτίμηση της ευθείας της παλινδρόμησης. Στην πραγματικότητα, τα δύο αυτά άτομα, αντιπροσωπεύουν επιτυχώς τόσο τον εαυτό τους όσο και εκείνους που δεν παρατηρήθηκαν, με αποτέλεσμα την διόρθωση του σφάλματος της προηγούμενης ανάλυσης.

2.4.3 Διόρθωση μέσω στάθμισης ομάδων

Σε κάθε πληθυσμιακή μελέτη, τα άτομα του δείγματος συνοδεύονται από ένα αντίστοιχο βάρος (base weight) w_{Bi}. Αυτό συνήθως είναι η αντίστροφη πιθανότητα επιλογής του ατόμου στο δείγμα. Αν θεωρήσουμε την πιθανότητα ένα άτομο i το οποίο έχει επιλεγεί στο δείγμα να ανταποκριθεί τότε έχουμε:

$$\begin{aligned} P(\text{το άτομο i να επιλεχθεί και να ανταποκριθεί}) = P(\text{το άτομο i να επιλεχθεί}) \cdot \\ P(\text{το άτομο i να ανταποκριθεί} \mid \text{ το άτομο i έχει επιλεχθεί}) = \\ = \pi_{Bi} \cdot \pi_{Ri} \end{aligned}$$

όπου π_{Bi} η πιθανότητα επιλογής στο δείγμα και π_{Ri} η πιθανότητα ανταπόκρισης δεδομένου ότι το άτομο έχει επιλεγεί. Αν η πιθανότητα π_{Ri} εκτιμηθεί, μπορεί να χρησιμοποιηθεί για να διορθώσει το σφάλμα λόγω μη ανταπόκρισης, σταθμίζοντας τους ανταποκριθέντες με το βάρος $w_{NRi} = \pi_{Ri}^{-1}$. Έτσι χρησιμοποιούμε το βάρος $w_{Bi} \cdot w_{NRi}$ για να διορθώσουμε για άνιση πιθανότητα επιλογής στο δείγμα και για μη ανταπόκριση. Στην πραγματικότητα η πιθανότητα ανταπόκρισης είναι άγνωστη ποσότητα και πρέπει να εκτιμηθεί. Η μείωση του σφάλματος μη ανταπόκρισης θα είναι τόσο σημαντική όσο η εκτιμημένη πιθανότητα ανταπόκρισης είναι σωστή. Δύο προσεγγίσεις είναι διαθέσιμες για την εκτίμηση αυτής της πιθανότητας. Η πρώτη είναι μέσω ταξινόμησης των ατόμων σε ομάδες μέσα στις οποίες η πιθανότητα ανταπόκρισης παραμένει σταθερή. Η δεύτερη είναι μέσω μοντελοποίησης της πιθανότητας αυτής.

Δημιουργία ομάδων με βάση το δείγμα

Αυτή η κατηγορία μεθόδων βασίζεται στον διαχωρισμό του δείγματος σε ομάδες, με χρήση πληροφοριών που διαθέτει ο ερευνητής για το δείγμα ή τον πληθυσμό. Ανάλογα με τα διαθέσιμα δεδομένα γίνεται η επιλογή της μεθόδου. Σε όλες όμως το κοινό είναι ο διαχωρισμός σε ομάδες, με βάση τις οποίες, υπολογίζονται βάρη τα οποία στην συνέχεια θα χρησιμοποιηθούν στην στατιστική ανάλυση. Ουσιαστικά επαναστάθμίζοντας, θεωρούμε τους ανταποκριθέντες αντιπροσωπευτικούς των μη ανταποκριθέντων κάθε κατηγορίας. Θα πρέπει λοιπόν μέσα σε κάθε ομάδα να υπάρχει ικανός αριθμός αναποκριθέντων. Αν σε κάποια κατηγορία δεν υπάρχουν ανταποκριθέντες είναι αδύνατος ο υπολογισμός των βαρών.

Η πρώτη προσέγγιση είναι ο διαχωρισμός του δείγματος με βάση χαρακτηριστικά που είναι γνωστά τόσο για τους ανταποκριθέντες όσο και για τους μη ανταποκριθέντες. Απαραίτητη είναι λοιπόν η γνώση κάποιων βασικών δημογραφικών χαρακτηριστικών και για τους μη ανταποκριθέντες. Για παράδειγμα είναι πιθανό το ποσοστό ανταπόκρισης να διαφοροποιείται ανάλογα την ηλικία και το φύλο. Ταξινομώντας τα άτομα του δείγματος σε K κατηγορίες μπορούμε να υπολογίσουμε το ποσοστό ανταπόκρισης για την k ομάδα. Αυτό το ποσοστό ανταπόκρισης, στην συνέχεια χρησιμοποιείται ως εκτίμηση της πιθανότητας ανταπόκρισης. Το ποσοστό ανταπόκρισης υπολογίζεται ως εξής:

$$RR_{wk} = \frac{\sum_{i=1}^{n_{rk}} w_{Bi}}{\sum_{i=1}^{n_k} w_{Bi}}$$

Το άθροισμα στον αριθμητή είναι στους ανταποκριθέντες του δείγματος μέσα στην k ομάδα, ενώ το άθροισμα στον παρονομαστή είναι σε όλα τα άτομα μέσα στην k ομάδα, ανταποκριθέντες και μη. Έχοντας υπολογίσει το ποσοστό ανταπόκρισης χρησιμοποιούμε το $w_{NRi} = RR_{wk}^{-1}$ ως παράγοντα διόρθωσης σε κάθε ανταποκριθέντα της k ομάδας. Η διόρθωση αυτή είναι πιο αποτελεσματική όσο η πραγματική πιθανότητα ανταπόκρισης είναι ομοιογενής μέσα στις ομάδες αυτές. Αν ακόμα, η πιθανότητα είναι σταθερή, τότε το σφάλμα μη ανταπόκρισης μπορεί να εξαλειφθεί τελείως. Κρίσιμο σημείο είναι λοιπόν η επιλογή των μεταβλητών που ορίζουν τις ομάδες αυτές. Οι μεταβλητές αυτές είναι κατάλληλες όταν συσχετίζονται ισχυρά τόσο με την υπο μελέτη μεταβλητή όσο και με την πιθανότητα ανταπόκρισης. Όταν υπάρχουν πολλές υπό μελέτη μεταβλητές, όπως συμβαίνει σε μια πληθυσμιακή μελέτη, η εύρεση τέτοιων μεταβλητών δεν είναι εύκολη υπόθεση.

Δημιουργία ομάδων με βάση τον πληθυσμό

Σε μια δεύτερη προσέγγιση, πληροφορία σχετική με τον συνολικό πληθυσμό είναι απαραίτητη. Θα δούμε δύο μεθόδους. Η πρώτη μπορεί να χρησιμοποιηθεί όταν είναι γνωστή η από κοινού κατανομή στον πληθυσμό των μεταβλητών με βάση τον οποίο θα γίνει ο διαχωρισμός (π.χ. κατανομή του πληθυσμού ανά ηλικία και φύλλο). Στην περίπτωση που μόνο οι περιθώριες κατανομές είναι γνωστές, είναι δυνατή η χρήση ενός επαναληπτικού αλγορίθμου.

Για την παρουσίαση θα γίνει χρήση ενός υποθετικού παραδείγματος (Kalton et al. 1998). Οι βοηθητικές μεταβλητές Α και Β χωρίζονται σε 4 και 3 κατηγορίες αντίστοιχα. Θεωρούμε εδώ ότι είναι κατηγορικές. Είναι γνωστή η από κοινού κατανομή των Α και Β στον πληθυσμό (πίνακας 2.2), ενώ η σταθμισμένη κατανομή τους στο δείγμα φαίνεται στον πίνακα 2.1. Τονίζεται εδώ ότι στον πίνακα 2.1 τα κελιά περιέχουν το άθροισμα των βαρών των ατόμων κάθε ομάδας και όχι τον αριθμό των ατόμων αυτών.

	B1	B2	B3	Σύνολο
A1	20	40	40	100
A2	50	140	310	500
A3	100	50	50	200
A4	30	100	70	200
Σύνολο	200	330	470	1000

Πίνακας 2.1: Σταθμισμένη Δειγματική Κατανομή (υποθετικό παράδειγμα)

	B1	B2	B3	Σύνολο
A1	80	40	55	175
A2	60	150	340	550
A3	170	60	200	430
A4	55	165	125	345
Σύνολο	365	415	720	1500

Πίνακας 2.2: Πληθυσμιακή Κατανομή (υποθετικό παράδειγμα)

Σύμφωνα με την θεωρία των πληθυσμιακών μελετών, αναμένουμε οι δύο κατανομές να είναι ίδιες. Κάτι τέτοιο δεν ισχύει εδώ. Συγκρίνοντας τους πίνακες βλέπουμε ότι το σύνολο για το

δείγμα είναι 1000 ενώ για τον πληθυσμό είναι 1500. Αυτή η διαφορά οφείλεται στην μη ανταπόκριση. Εάν συγκρίνουμε και τα αντίστοιχα κελιά του δείγματος και του πληθυσμού, παρατηρούνται σημαντικές διαφορές. Κατά την βασική διαδικασία στάθμισης των κελιών, τα δειγματοληπτικά βάρη προσαρμόζονται έτσι ώστε το δείγμα να συμφωνεί με τον πληθυσμό κελί-ανά-κελί. Έτσι για το κελί A1-B1 η διόρθωση είναι 80/20=4, ενώ για το A1-B2 είναι 40/40=1. Στον πίνακα 2.3 παρουσιάζονται οι διορθώσεις για όλα τα κελιά.

	B1	B2	B3
A1	4.00	1.00	1.37
A2	1.20	1.07	1.10
A3	1.70	1.20	4.00
A4	1.83	1.65	1.79

Πίνακας 2.3: Διορθώσεις των κελιών

Αυτή η μέθοδος υποθέτει ότι οι ανταποκριθέντες σε κάθε κελί είναι αντιπροσωπευτικοί των μη ανταποκριθέντων του ίδιου κελιού. Κάνοντας εκτίμηση της πιθανότητας ανταπόκρισης (όπως στην περίπτωση της δημιουργίας ομάδων με βάση το δείγμα) αυτή η υπόθεση ικανοποιείται εάν κάθε άτομο σε ένα κελί έχει την ίδια πιθανότητα ανταπόκρισης δεδομένης της επιλογής στο δείγμα. Αυτό αντιστοιχεί στην υπόθεση περί τυχαίας έλλειψης δεδομένων (MAR). Αντιθέτως η μέθοδος στάθμισης που παρουσιάζεται εδώ δεν κάνει καμία τέτοια υπόθεση για την δομή της πιθανότητας ανταπόκρισης.

Ένα μειονέκτημα της μεθόδου είναι ότι μπορεί να οδηγήσει σε μεγάλες διακυμάνσεις των βαρών, αυξάνοντας έτσι την διακύμανση των εκτιμήσεων της μελέτης. Ιδιαίτερα εάν το μέγεθος των κελιών είναι μικρό (π.χ. όταν ο αριθμός των κελιών είναι μεγάλος) αυτό μπορεί να αποτελεί σημαντικό πρόβλημα, αφού μικρο μέγεθος δείγματος μπορεί να οδηγήσει σε αστάθεια στην διόρθωση. Γι αυτό το λόγο χρειάζεται είτε να γίνει περικοπή των βαρών (π.χ. στο 99ο ποσοστημόριο)είτε να γίνει χρήση κάποιας εναλλακτικής μεθόδου.

Raking:

Σε αντίθεση με την μέθοδο στάθμισης ανά κελί, κατά την οποία η από κοινού κατανομή των βοηθητικών μεταβλητών στο δείγμα αλλάζει ώστε να συμφωνεί με αυτήν του πληθυσμού, η μέθοδος raking βασίζεται μόνο στις περιθώριες κατανομές των μεταβλητών. Η μέθοδος raking είναι ένας επαναληπτικός αλγόριθμος.

Ο αλγόριθμος περιγράφεται παρακάτω:
Αν θεωρήσουμε το παράδειγμα με τις δύο μεταβλητές, με τις περιθώριες κατανομές των μεταβλητών να είναι γνωστές για τον πληθυσμό, και η από κοινού κατανομή μόνο για το δείγμα. Ξεκινώντας από τις γραμμές του πίνακα του δείγματος και παίρνοντας με την σειρά κάθε γραμμή, πολλαπλασιάζουμε το κάθε κελί στην γραμμή με τον λόγο του πληθυσμιακού συνόλου προς το σταθμισμένο σύνολο του δείγματος για εκείνη την ομάδα. Έτσι τα σταθμισμένα σύνολα των γραμμών συμφωνούν με τα σύνολα του πληθυσμού για αυτή την μεταβλητή. Όμως, κατά πάσα πιθανότητα, τα καινούρια σύνολα ανά στήλη δεν θα συμφωνούν με εκείνα του πληθυσμού. Επαναλαμβάνουμε την διαδικασία και για τις στήλες του πίνακα. Πάλι όμως τα σύνολα των γραμμών θα αλλάξουν και πιθανόν να μην συμφωνούν με τον πληθυσμό. Η διαδικασία συνεχίζεται, εναλλάσσοντας μεταξύ γραμμών και στηλών και η ταυτόχρονη συμφωνία γραμμών και στηλών με εκείνες του πληθυσμού αναμένεται να επιτευχθεί μετά από μερικές επαναλήψεις. Το αποτέλεσμα είναι ένας πίνακας για τον πληθυσμό που αντανακλά την σχέση μεταξύ των δύο μεταβλητών στο δείγμα (Battaglia et al. 2013). Στο παράδειγμα που

χρησιμοποιήσαμε παραπάνω με την εφαρμογή του αλγόριθμου παίρνουμε τα αποτελέσματα του πίνακα 2.4.

	B1	B2	B3
A1	1.81	1.45	2.02
A2	1.08	0.87	1.21
A3	2.20	1.76	2.45
A4	1.83	1.47	2.04

Πίνακας 2.4: Διορθώσεις που προκύπτουν από τη μέθοδο Raking

Η μέθοδος raking υποθέτει ότι οι πιθανότητες ανταπόκρισης είναι ίσες μεταξύ των ατόμων σε κάθε κελί , όπως και η cell weighting μέθοδος. Επιπλέον όμως υποθέτει ότι η πιθανότητα ανταπόκρισης π_{ij} για το κελί (i, j) είναι της μορφής $\pi_{ij} = \alpha_i \cdot \beta_j$ (Kalton & Maligalig 1991). Δηλαδή οι μεταβλητές A και B δεν αλληλεπιδρούν στην πιθανότητα ανταπόκρισης. Αυτή η τελευταία υπόθεση οδηγεί σε μείωση της μεταβλητότητας των βαρών αλλά και σε σημαντική διαφοροποίηση των αποτελεσμάτων σε σύγκριση με την στάθμιση ανά κελί. Σε τέτοιες περιπτώσεις είναι καλό να εξετάζεται προσεκτικά η χρήση αυτής της μεθόδου.

2.4.4 Μοντελοποίηση της Πιθανότητας Ανταπόκρισης

Όταν έχει παρατηρηθεί αρκετή πληροφορία τόσο για τους ανταποκριθέντες όσο και για τους μη ανταποκριθέντες, οι προηγούμενες μέθοδοι δεν είναι εύκολα εφαρμόσιμες, λόγω του μεγάλου αριθμού των κελιών που δημιουργούνται .Αυτό που προτείνεται είναι η μοντελοποίηση της πιθανότητας ανταπόκρισης, συνήθως μέσω λογιστικής παλινδρόμησης. Αρχικά θεωρούμε έναν δείκτη R τέτοιο ώστε $R_i = 1$ εάν η παρατήρηση είναι πλήρης και $R_i = 0$ διαφορετικά. Η πιθανότητα $P(R = 1)$ είναι η πιθανότητα η παρατήρηση να είναι πλήρης (πιθανότητα ανταπόκρισης). Για να μοντελοποιήσουμε την πιθανότητα αυτή θα θέλαμε να βρούμε ένα σύνολο μεταβλητών $\mathbf{H}$ τέτοιο ώστε να ισχύει:

$$P(R = 1 \mid \mathbf{X}, Y, \mathbf{H}) = P(R = 1 \mid \mathbf{H}) \tag{2.4.1}$$

με $\mathbf{X}$ το σύνολο των ανεξάρτητων μεταβλητών της ανάλυσης και Υ την μεταβλητή απόκρισης.

Επιπλέον προϋπόθεση είναι να μοντελοποιήσουμε σωστά την σχέση μεταξύ R και $\mathbf{H}$. Στην πράξη χρησιμοποιούνται κυρίως μοντέλα λογιστικής παλινδρόμησης.

2.4.5 Προσδιορισμός του μοντέλου

Το πρώτο που μας απασχολεί είναι η επιλογή των ανεξάρτητων μεταβλητών. Οι μεταβλητές που συσχετίζονται με την πιθανότητα $R = 1$ είναι λογικό να συμπεριληφθούν στο μοντέλο ως ανεξάρτητες μεταβλητές. Ωστόσο, αυτό δεν είναι απαραίτητο. Για παράδειγμα οι μεταβλητές εκείνες οι οποίες προβλέπουν την πιθανότητα $R = 1$ ανεξάρτητα από τις $\mathbf{X}$ και Y, είναι καλό να μην περιλαμβάνονται. Εάν η εξίσωση (2.4.1) ισχύει και $\mathbf{H} = (H_a, H_b)$ όπου η H_b είναι ανεξάρτητη του $\mathbf{X}$ και Y δοθέντος H_a, τότε η εξίσωση (2.4.1) θα εξακολουθεί να ισχύει ακόμα και αν το H αντικατασταθεί με το H_a, δηλαδή το H_b δεν χρειάζεται. Εάν, παρόλα αυτά, το συμπεριλάβουμε, τότε δύο άτομα με ίδιο H_a αλλά διαφορετικό H_b θα λάβουν διαφορετικά βάρη, ενώ θα έπρεπε να λάβουν ίδια. Αυτό θα αυξήσει μόνο την μεταβλητότητα των βαρών.

Αντίθετα, μεταβλητές που δεν σχετίζονται με την πιθανότητα ανταπόκρισης αλλά σχετίζονται και τα $\mathbf{X}$ και Y πρέπει να περιληφθούν.

Η προσθήκη όμως όλο και περισσότερων μεταβλητών, σε ένα πεπερασμένο σε μέγεθος δείγμα, εν τέλει θα οδηγήσει σε προβλεπόμενη πιθανότητα ίση με μηδέν για τουλάχιστον μια ελλείπουσα παρατήρηση. Αυτό σημαίνει ότι η μεταβλητή **H** είναι τόσο πληροφοριακή που δεν υπάρχουν παρόμοιες παρατηρήσεις, έτσι το άτομο για το οποίο δεν έχουμε πληροφορία, δεν μπορεί να αντιπροσωπευθεί από κανένα άτομο στο δείγμα.

Συμπερασματικά, το μοντέλο θα πρέπει να περιέχει αρκετές μεταβλητές ώστε η (2.4.1) να ισχύει, άλλα όχι απαραίτητα όλες της μεταβλητές που σχετίζονται με την πιθανότητα ανταπόκρισης. Η προσθήκη επιπλέον μεταβλητών μπορεί να αυξήσει την ακρίβεια των αποτελεσμάτων (Seaman & White 2013).

Όλα τα προηγούμενα ισχύουν υπό την προϋπόθεση ότι η σχέση μεταξύ **H** και R είναι γνωστό πως να μοντελοποιηθεί σωστά. Το μοντέλο λογιστικής παλινδρόμησης χρησιμοποιείται ως επί τω πλείστον για λόγους ευκολίας. Αυτό δεν σημαίνει ότι είναι πάντα σωστό. Επιπλέον, μετασχηματισμοί και όροι αλληλεπίδρασης μπορεί να χρειαστούν για να το κάνουν σωστό.

Ένα ακόμα πρόβλημα το οποίο μπορεί να προκύψει είναι αυτό των ασταθών βαρών. Κάποια άτομα μπορεί να έχουν μικρή προβλεπόμενη πιθανότητα, άρα μεγάλα βάρη, όχι επειδή έχουν όντως μικρή πιθανότητα, αλλά επειδή το μοντέλο δεν είναι σωστά προσδιορισμένο. Όταν τέτοια προβλήματα προκύπτουν είναι εύλογο να αναρωτηθεί κανείς εάν το γεγονός αυτό οφείλεται σε μεγάλη προβλεπτική ικανότητα των ανεξάρτητων μεταβλητών ή σε κακό προσδιορισμό του μοντέλου.

Για τον έλεγχο της καλής προσαρμογής του μοντέλου, προτείνεται ο έλεγχος Hosmer–Lemeshow (Hosmer Jr & Lemeshow 2004). Ένας άλλος τρόπος που προτείνεται είναι ο έλεγχος του Hinkley (Hinkley 1985). Με βάση αυτή την προσέγγιση, τα τετράγωνα των προβλεπόμενων πιθανοτήτων, εισάγονται σε ένα νέο μοντέλο λογιστικής παλινδρόμησης. Εάν αυτός ο επιπλέον όρος είναι στατιστικά σημαντικός, υπάρχει ένδειξη ότι το μοντέλο δεν κάνει καλή εφαρμογή στα δεδομένα στις ουρές και γι αυτό τα μεγάλα βάρη ίσως δεν είναι σωστά εκτιμημένα.

Για την αντιμετώπιση των μεγάλων βαρών, οι προτεινόμενες μέθοδοι περιλαμβάνουν την περικοπή των βαρών, ημι-παραμετρική μοντελοποίηση με λογιστική παλινδρόμηση και άλλου είδους μοντελοποιήσεις.

Για να γίνει περικοπή των μεγάλων βαρών επιλέγεται ένα μέγιστο βάρος και όλα τα μεγαλύτερα από το μέγιστο βάρη τίθενται ίσα με αυτό. Όμως η περικοπή των βαρών μπορεί να ξανά εισάγει ένα μέρος του σφάλματος που η στάθμιση αντίστροφης πιθανότητας έχει εξαλείψει, εάν τα μεγάλα βάρη οφείλονται στην μεγάλη προβλεπτική ικανότητα των εξαρτημένων μεταβλητών. Στην περίπτωση που τα μεγάλα βάρη οφείλονται σε κακό προσδιορισμό του μοντέλου, η περικοπή τους είναι λογική. Η επιλογή της μέγιστης τιμής γίνεται εντελώς αυθαίρετα. Είναι λοιπόν σημαντικό να επιβεβαιωθεί ότι οι τελικές εκτιμήσεις δεν είναι ιδιαίτερα ευαίσθητες σε αλλαγές της μέγιστης τιμής.

Ημιπαραμετρικά μοντέλα μπορούν επίσης να χρησιμοποιηθούν (Wang et al. 1997) για την μοντελοποίηση, όμως αυτή η προσέγγιση περιορίζεται σε περιπτώσεις με μικρό αριθμό ανεξάρτητων μεταβλητών. Μια διαφορετική πρόταση είναι η χρήση της robit regression (Kang & Schafer 2007, Liu 2004). Η robit regression αντικαθιστά την συνδετική συνάρτηση της λογιστικής με την συνάρτηση κατανομής μιας t με ν βαθμούς ελευθερίας. Η t κατανομή έχει παχύτερες ουρές (heavy tailed) κάνοντας την παλινδρόμηση πιο ανθεκτική σε ακραίες παρατηρήσεις και λιγότερο επιρρεπή στην παραγωγή μικρών βαρών όταν το μοντέλο δεν έχει προσδιοριστεί σωστά.

Μια πρόταση για τον προσδιορισμό του μοντέλου περιλαμβάνει τα παρακάτω βήματα (Seaman & White 2013):

Πρώτο, ο προσδιορισμός μεταβλητών με καλή προβλεπτική ικανότητα. Καλό θα ήταν να αφαιρεθούν εκείνες οι οποίες είναι ανεξάρτητες των X και Υ και να προστεθούν εκείνες που είναι ισχυρά συσχετιζόμενες με το Υ. Αν τα βάρη πρόκειται να χρησιμοποιηθούν για πολλές αναλύσεις, δηλαδή πολλές διαφορετικές μεταβλητές απόκρισης Υ, το τελευταίο δεν είναι δυνατόν να γίνει.

Δεύτερο, η εξέταση της κατανομής των συνεχών ανεξάρτητων μεταβλητών. Πιθανόν να χρειαστεί κάποιος μετασχηματισμός σε αυτές.

Τρίτο, η εφαρμογή του μοντέλου με χρήση όλων των μεταβλητών του πρώτου βήματος. Αν είναι δυνατή η εφαρμογή robit regression καλό είναι να γίνει. Αν υπάρχουν πολλές υποψήφιες μεταβλητές τότε μπορεί να γίνει χρήση κάποιας αυτοματοποιημένης μεθόδου επιλογής π.χ. forward selection. Σε τέτοια περίπτωση καλό είναι να εξαναγκάσουμε κάποιες μεταβλητές να εισαχθούν στο μοντέλο π.χ. ηλικία και φύλο. Επιπλέον, σε αυτό το βήμα, μπορεί να γίνει προσθήκη όρων αλληλεπίδρασης.

Τέταρτο, ο έλεγχος προσαρμογής του μοντέλου με το τεστ Hosmer–Lemeshow (Hosmer Jr & Lemeshow 2004) και/ή την μέθοδο του Hinkley (Hinkley 1985).

Πέμπτο, ο έλεγχος της κατανομής των βαρών και η εξέταση ύπαρξης μηδενικών προβλεπόμενων πιθανοτήτων. Σε περίπτωση ύπαρξης μηδενικών θα πρέπει να αφαιρεθούν κάποιες ανεξάρτητες μεταβλητές ή να συγχωνευτούν κατηγορίες κατηγορικών μεταβλητών. Υπάρχει, βεβαίως, και η περίπτωση η πραγματική πιθανότητα για κάποια άτομα να είναι μηδέν ή σχεδόν μηδέν, οπότε αυτή η μεθοδολογία δεν μπορεί να εφαρμοστεί στα δεδομένα. Σε αυτό το βήμα πρέπει να γίνει έλεγχος για την σταθερότητα των βαρών . Σε περίπτωση μη σταθερότητας επιπλέον τροποποιήσεις μπορεί να γίνουν στο μοντέλο π.χ. μετασχηματισμοί συνεχών μεταβλητών ή επιπλέον όροι αλληλεπίδρασης ή αλλαγή της μοντελοποίησης. Στην περίπτωση που γίνει περικοπή των βαρών πρέπει να γίνει και έλεγχος ευαισθησίας για την επιλογή της μέγιστης τιμής. Βέβαια η αστάθεια των βαρών μπορεί και να αντανακλά πραγματικές διαφορές μεταξύ των ανταποκριθέντων και μη ανταποκριθέντων επομένως ίσως να πρέπει να γίνουν αποδεκτές οι μεγάλες τυπικές αποκλίσεις ως έκφραση της πραγματικής αβεβαιότητας.

2.4.6 Επαυξημένη Στάθμιση με Αντίστροφη Πιθανότητα Διπλά Ανθεκτική Εκτιμήτρια

Οι μέθοδοι Στάθμισης με Αντίστροφη Πιθανότητα μπορούν να χρησιμοποιηθούν και στην περίπτωση μη ανταπόκρισης ατόμου αλλά και ερώτησης. Για την αντιμετώπιση της δεύτερης περίπτωσης έχουν προταθεί μέθοδοι Επαυξημένης Στάθμισης με Αντίστροφη Πιθανότητα (Augmented IPW) (Seaman & White 2013, Bang & Robins 2005), ή αλλιώς διπλά ανθεκτικές εκτιμήτριες.

Για την εφαρμογή μιας διπλά ανθεκτικής εκτίμησης, δύο μοντέλα ορίζονται:
1) Το μοντέλο ανταπόκρισης το οποίο απαιτεί τον προσδιορισμό ενός μοντέλου το οποίο περιγράφει τον άγνωστο μηχανισμό μη ανταπόκρισης,
2) Το μοντέλο έκβασης, το οποίο περιγράφει την κατανομή της υπό μελέτη μεταβλητής.

Μια εκτιμήτρια ονομάζεται διπλά ανθεκτική (Doubly Robust) εάν παραμένει ασυμπτωτικά αμερόληπτη και συνεπής ακόμα και εάν ένα από τα δύο μοντέλα (μη ανταπόκρισης ή έκβασης) δεν είναι σωστά ορισμένο. Η διαδικασία αυτή προσφέρει κάποια προστασία σε περίπτωση λάθως προσδιορισμού ενός μοντέλου (Kim & Haziza 2010).

Ας υποθέσουμε ότι έχουμε n ανεξάρτητες παρατηρήσεις από μία μεταβλητή Υ, $y_1, y_2, ..., y_n$ και ενδιαφερόμαστε να εκτιμήσουμε την μέση τιμή της μεταβλητής, $E(Y) = \theta$. Απουσία μη ανταπόκρισης η εκτίμηση γίνεται από τον δειγματικό μέσο.

$$\hat{\theta}_n = \sum_{i=1}^{n} w_i y_i \tag{2.4.2}$$

όπου $w_i = 1/n$ τα δειγματοληπτικά βάρη. Η μέθοδος επεκτείνεται και στην περίπτωση που τα βάρη είναι διαφορετικά μεταξύ τους. Έστω $\mathbf{x}$ ένα διάνυσμα επεξηγηματικών μεταβλητών και R_i μια δείκτρια μεταβλητή τέτοια ώστε $R_i = 1$ εάν η y_i έχει παρατηρηθεί και $R_i = 0$ διαφορετικά. Έτσι παρατηρούμε τα $(\mathbf{x}_i, y_i)$ όταν $R_i = 0$ ενώ μόνο τα $\mathbf{x}_i$ όταν $R_i = 1$.

Για την εκτίμηση του θ ορίζουμε πρώτα ένα μοντέλο για την δεσμευμένη κατανομή των y_i δοθέντων των $\mathbf{x}_i$. Συγκεκριμένα εάν ενδιαφερόμαστε μόνο για τη μέση τιμή του Υ ορίζουμε το

παρακάτω μοντέλο:

$$E(y_i \mid \mathbf{x}_i, R_i = 0) = m(\mathbf{x}_i; \beta_0), \tag{2.4.3}$$

όπου $m(\mathbf{x}_i, \beta)$ μια συνεχής, παραγωγίσιμη συνάρτηση του β. Το μοντέλο (2.4.3) ονομάζουμε μοντέλο έκβασης και ουσιαστικά αποτελεί ένα μοντέλο αντικατάστασης. Η εκτίμηση του θ δίνεται από την παρακάτω σχέση:

$$\hat{\theta}_1 = \sum_{i=1}^{n} w_i\{R_i y_i + (1 - R_i)m(\mathbf{x}_i, \hat{\beta})\}, \tag{2.4.4}$$

όπου $\hat{\beta}$ μια συνεπής εκτίμηση της πραγματικής τιμής της παραμέτρου β_0. Ουσιαστικά η εκτιμήτρια (2.4.4) χρησιμοποιεί τις παρατηρηθείσες τιμές της μεταβλητής U όπου αυτή έχει παρατηρηθεί ($R_i = 1$), ενώ τις προβλεπόμενες από το μοντέλο (2.4.3) όταν αυτή δεν έχει παρατηρηθεί ($R_i = 0$). Υπολογίζει έτσι την μέση τιμή χρησιμοποιώντας όλο το δείγμα.

Υποθέτουμε επιπλέον, ότι η πιθανότητα ανταπόκρισης στην υπό μελέτη μεταβλητή, την οποία συμβολίζουμε με $p_i = Pr(R_i = 1 \mid i)$, περιγράφεται από ένα μοντέλο λογιστικής παλινδρόμησης:

$$p_i = p_i(\phi_0) = \frac{exp(\phi_0' x_i)}{1 + exp(\phi_0' x_i)} \tag{2.4.5}$$

για κάποιο ϕ_0. Το μοντέλο (2.4.5) καλείται μοντέλο μη ανταπόκρισης.

Ορίζουμε λοιπόν τον εκτιμητή :

$$\hat{\theta}_{tp} = \sum_{i=1}^{n} w_i\{m(x_i, \hat{\beta}) + \frac{R_i}{p_i}(y_i - m(\mathbf{x}_i, \hat{\beta}))\}, \tag{2.4.6}$$

ο οποίος ισοδύναμα γράφεται:

$$\hat{\theta}_{tp} = \hat{\theta}_n + \sum_{i=1}^{n} w_i\{(\frac{R_i}{p_i} - 1)(y_i - m(\mathbf{x}_i, \hat{\beta}))\}, \tag{2.4.7}$$

και είναι αμερόληπτος για το θ κάτω από το μοντέλο μη ανταπόκρισης (Cochran 1997) ανεξάρτητα από το εάν το μοντέλο έκβασης της σχέσης (2.4.3) ισχύει. Επίσης, όταν το μοντέλο ανταπόκρισης δεν είναι σωστό, παραμένει αμερόληπτος εάν ισχύει η (2.4.3) και ο μηχανισμός είναι MAR . Έτσι ο εκτιμητής $\hat{\theta}_{tp}$ έχει την διπλά ανθεκτική ιδιότητα.

Όταν η πιθανότητα ανταπόκρισης εκτιμάται αντί να είναι γνωστή, θεωρούμε τον εκτιμητή:

$$\hat{\theta}_{DR}(\hat{\beta}, \hat{\phi}) = \hat{\theta}_n + \sum_{i=1}^{n} w_i\{(\frac{R_i}{p_i(\hat{\phi})} - 1)(y_i - m(\mathbf{x}_i, \hat{\beta}))\}, \tag{2.4.8}$$

Η διπλά ανθεκτική ιδιότητα εξακολουθεί να ισχύει και σε αυτή την περίπτωση (Scharfstein et al. 1999). Είναι σημαντικό να τονίσουμε σε αυτό το σημείο ότι η σχέση (2.4.8) ορίζει μια κατηγορία εκτιμητριών καθώς διαφορετική μέθοδος εκτίμησης των $(\hat{\beta}, \hat{\phi})$ οδηγεί σε διαφορετική εκτιμήτρια (Cao et al. 2009).

2.4.7 Εφαρμογή της μεθόδου για μη ανταπόκριση σε επίπεδο ατόμου. Το παράδειγμα των:

NHANES

Εδώ θα γίνει περιγραφή του πως εφαρμόζονται οι παραπάνω μέθοδοι για μη ανταπόκριση ατόμου από την μελέτη NHANES. Η NHANES (National Health and Nutrition Examination Survey)

είναι μια μελέτη σχετικά με την υγεία η οποία διεξάγεται από το CDC (Center for Disease Control and Prevention) και το NCHS (National Center for Health Statistics) και συγκεντρώνει πληροφορίες σχετικά με την υγεία και την διατροφή του πληθυσμού των ΗΠΑ, εκτιμά τον επιπολασμό διάφορων ασθενειών και παθήσεων και παρέχει πληροφορίες για την οργάνωση της στρατηγικής της υγείας.

Στο δημοσιευμένο πρωτόκολλο περιγράφονται αναλυτικά οι μέθοδοι με τις οποίες πραγματοποιείται η διόρθωση για την μη ανταπόκριση ατόμου (Mirel LB 2013).

Η διόρθωση για την μη ανταπόκριση γίνεται ξεχωριστά για τις τρείς φάσεις της μελέτης, (screening, interview, examination). Οι ομάδες στάθμισης δημιουργούνται με χρήση μεταβλητών γνωστών και για τους ανταποκριθέντες και για τους μη ανταποκριθέντες. Για την πρώτη φάση της μελέτης χρησιμοποιούνται μεταβλητές σχετικές με την τοποθεσία, ενώ για τις δύο επόμενες χρησιμοποιείται το Chi-squared Automatic Interaction Detector (CHAID) για τον εντοπισμό μεταβλητών με υψηλή συσχέτιση με την πιθανότητα ανταπόκρισης. Οι μεταβλητές αυτές είναι γνωστές από τις προηγούμενες φάσεις στις οποίες υπήρχε ανταπόκριση.

Έτσι οι παράγοντες διόρθωσης σχηματίζονται με την παρακάτω σχέση

$$f_{i(NR)} = \frac{\sum_{i=1}^{n_{a_s}} w_{i(base)}}{\sum_{i=1}^{n_{a_r}} w_{i(base)}}$$

όπου $w_{i(base)}$ το αρχικό βάρος για τον i συμμετέχοντα που έχει συμπεριληφθεί στο δείγμα στο α-οστό κελί, n_{a_s} το συνολικό δείγμα στο α-οστό κελί και n_{a_r} οι ανταποκριθέντες στο α-οστό κελί. Η άθροιση έγινε ξεχωριστά για κάθε κελί. Έτσι τα διορθωμένα βάρη υπολογίστηκαν ως εξής :

$$w_{i(NR)} = w_{i(base)} f_{i(NR)}$$

Στη συνέχεια έγινε περικοπή (trimming) των βαρών αφού μεγάλα βάρη μπορούν να διογκώσουν την διασπορά των εκτιμήσεων. Η περικοπή έγινε στα τμήματα του δείγματος που ήταν σχεδιασμένα να είναι αυτό-σταθμιζόμενα. Αυτά τα τμήματα ήταν σχετικά με την φυλή (Ισπανική) και καταγωγή-εισόδημα-ηλικία-φύλο. Επειδή κάποιες ομάδες έχουν υπέρ-δειγματοληπτηθεί, τα βάρη μέσα σε κάποιους τομείς μπορεί να ποικίλουν αρκετά. Γι αυτό το λόγο τα όρια της περικοπής εξαρτώνται από το πόσο υπέρ-δειγματοληψία έγινε στους τομείς αυτούς. Όταν τα βάρη που πρέπει να περικοπούν εντοπιστούν, τα βάρη αυτά που δεν θα περικοπούν πρέπει να προσαρμοστούν ώστε το άθροισμα των βαρών πριν και μετά την περικοπή να μην αλλάξει. Οι παράγοντες περικοπής υπολογίζονται ως εξής:

$$f_{i(TR)} = \frac{\sum_{i=1}^{n_b} t_i}{\sum_{i=1}^{n_b} w_{i(base)} f_{i(NR)}}$$

όπου n_b το μέγεθος του b-οστού φυλή-(Ισπανική) καταγωγή-εισόδημα-ηλικία-φύλο δειγματικού τομέα, και t_i το γινόμενο $w_{i(base)} f_{i(NR)}$, δεδομένου ότι το γινόμενο αυτό δεν υπερβαίνει το προκαθορισμένο όριο διαφορετικά ισούται με το όριο αυτό. Τα περικεκομένα βάρη υπολογίζονται τότε ως εξής:

$$w_{i(TR)} = w_{i(NR)} f_{i(TR)}$$

Το τελευταίο βήμα της διαδικασίας είναι η εκ των υστέρων στρωματοποιηση με χρήση γνωστών τιμών του πληθυσμού. Η διαδικασία αυτή χρησιμοποιείται για να διορθώσει για μη κάλυψη, διορθώνει όμως και για κατάλοιπα σφάλματα από μη ανταπόκριση. Ένα επιπλέον όφελος είναι η μείωση της διασποράς των εκτιμήσεων. Μεγάλες κλάσεις δημιουργούνται (εκ των υστέρων στρώματα) με χρήση βοηθητικής πληροφορίας και ένας διορθωτικός παράγοντας εφαρμόζεται σε όλες τις μονάδες σε ένα δεδομένο εκ των υστέρων στρώμα. Ουσιαστικά η τεχνική αυτή αυξάνει τα σταθμισμένα σύνολα

στο επίπεδο του πληθυσμού. Οι παράγοντες της εκ των υστέρων στρωματοποίησης υπολογίζονται ως εξής:

$$f_{i(PS)} = \frac{N_c}{\sum_{i=1}^{n_c} w_{i(TR)}}$$

όπου N_c το σύνολο του πληθυσμού και n_c το μέγεθος του δείγματος σε ορισμένο εκ των υστέρων στρώμα. Έτσι τα βάρη της εκ των υστέρων στρωματοποίησης υπολογίζονται ως εξής:

$$w_{i(PS)} = w_{i(NR)} f_{i(PS)}$$

Τα τελικά βάρη για κάθε επιλεγμένο άτομο για κάθε φάση της μελέτης υπολογίστηκαν ως το γινόμενο του αρχικού βάρους, της διόρθωσης για μη ανταπόκριση, της περικοπής και της εκ των υστέρων στρωματοποίησης. Δηλαδή:

$$w_i = w_{i(base)} f_{i(NR)} f_{i(TR)} f_{i(PS)}$$

Πιο συγκεκριμένα για την πρώτη φάση της μελέτης (screening) έγινε ο εξής υπολογισμός:

$$w_{i(S)} = w_{i(base)} f_{i(NR,S)} f_{i(TR,S)} f_{i(PS,S)}$$

Ενώ για τους ανταποκριθέντες της δεύτερης φάσης (συνέντευξη) η διαδικασία επαναλήφθηκε με χρήση των επιπλέον δεδομένων που είχαν συλλεχθεί για τους ανταποκριθέντες της πρώτης φάσης ο οποίοι δεν ανταποκρίθηκαν στην συνέντευξη. Έτσι με $f_{i(NR,I)}, f_{i(TR,I)}$, και $f_{i(PS,I)}$ συμβολίζονται οι παράγοντες διόρθωσης μη-ανταπόκρισης, περικοπής και εκ των υστέρων στρωματοποίησης αντίστοιχα, που εφαρμόστηκαν στους ανταποκριθέντες της συνέντευξης. Έγινε λοιπόν ο εξής υπολογισμός:

$$w_{i(I)} = w_{i(base)} f_{i(NR,S)} f_{i(TR,S)} f_{i(PS,S)} f_{i(NR,I)} f_{i(TR,I)} f_{i(PS,I)}$$

Ομοίως για τους ανταποκριθέντες της τρίτης φάσης (εξετάσεις υγείας) έγινε χρήση των δεδομένων που είχαν συλλεχθεί για τους ανταποκριθέντες της δεύτερης φάσης ο οποίοι δεν ανταποκρίθηκαν στην τρίτη φάση. Έτσι με $f_{i(NR,E)}, f_{i(TR,E)}$, και $f_{i(PS,E)}$ συμβολίζονται οι παράγοντες διόρθωσης μη-ανταπόκρισης, περικοπής και εκ των υστέρων στρωματοποίησης αντίστοιχα, που εφαρμόστηκαν στους ανταποκριθέντες των εξετάσεων υγείας. Έγινε λοιπόν ο εξής υπολογισμός:

$$w_{i(E)} = w_{i(base)} f_{i(NR,S)} f_{i(TR,S)} f_{i(PS,S)} f_{i(NR,I)} f_{i(TR,I)} f_{i(PS,I)} f_{i(NR,E)} f_{i(TR,E)} f_{i(PS,E)}$$

India Demographic Health Survey

Στα πλαίσια μιας ανάλυσης για την εκτίμηση του ποσοστού των Ινδών αντρών που είχαν σεξουαλική συνεύρεση επί πληρωμή, της μελέτης India Demographic Health Survey, έγινε διόρθωση για μη ανταπόκριση ερώτησης, χρησιμοποιώντας μεθόδους στάθμισης (Wirth et al. 2010). Χρησιμοποιήθηκε επιπλέον και η μέθοδος της επαυξημένης στάθμισης με αντίστροφη πιθανότητα (Augmented IPW) η οποία αναφέρεται και ως διπλά ανθεκτική (doubly robust). Στην συνέχεια έγινε σύγκριση των αποτελεσμάτων της αναλύσης που λαμβάνει υπόψιν την μη ανταπόκριση σε σχέση με εκείνη που την αγνοεί, ενώ έγινε και σύγκριση των αποτελεσμάτων των διαφορετικών μεθόδων που χρησιμοποιήθηκαν.

Αρχικά έγινε εφαρμογή ενός μοντέλου λογιστικής παλινδρόμησης για την πιθανότητα να έχει απαντηθεί η ερώτηση σχετικά με την σεξουαλική συνεύρεση επί πληρωμή την οποία θα συμβολίζουμε με $R = 1$, δεδομένου ενός συνόλου ανεξάρτητων μεταβλητών. Για την επιλογή των μεταβλητών αυτών αρχικά συγκεντρώθηκαν δημογραφικές και βιβλιογραφικές πληροφορίες σχετικά με το ποιες

μεταβλητές μπορεί να σχετίζονται με την $P(R = 1)$. Συμπεριλήφθηκαν όλοι οι όροι αλληλεπίδρασης και έγινε έλεγχος για γραμμικότητα με τεστ πηλίκου πιθανοφάνειας για κάποιες από αυτές. Στην συνέχεια εφαρμόστηκε μια αυτοματοποιημένη μέθοδος επιλογής (stepwise forward selection). Τελικά επιλέχθηκαν 55 μεταβλητές και όροι αλληλεπίδρασης. Συμβολίζοντας το σύνολο των ανεξάρτητων αυτών μεταβλητών ως $\bar{M}$ μοντελοποιήθηκε η $P(R = 1 \mid \bar{M})$ ως εξής:

$$\log(\frac{P(R = 1 \mid \bar{M}; a)}{1 - P(R = 1 \mid \bar{M}; a)}) = a_0 + \sum_{j=1}^{55} a_j M_j$$

Οι αντίστροφες προβλεπόμενες από το μοντέλο πιθανότητες πολλαπλασιάστηκαν με τα αρχικά βάρη της μελέτης για τον σχηματισμό των τελικών βαρών και έγινε η εκτίμηση του μέσου ως εξής:

$$\bar{\mu}_{IPW} = \frac{\sum_{i=1}^{N} W *_{i,s} Y_i}{\sum_{i=1}^{N} W *_{i,s}}$$

όπου $W *_{i,s}$ το τελικό βάρος για το i άτομο και Y_i η παρατηρούμενη τιμή της μεταβλητής.

Στην συνέχεια εφαρμόστηκε ένα δεύτερο μοντέλο λογιστικής παλινδρόμησης, αυτή τη φορά , με εξαρτημένη μεταβλητή το αν έχει παρατηρηθεί συνεύρεση ή όχι. Αυτή η μέθοδος αναφέρεται ως outcome regression. Η μεταβλητή Υ=1 εάν έχει παρατηρηθεί συνεύρεση επί πληρωμή και Υ=0 διαφορετικά. Επιλέχθηκαν 50 ανεξάρτητες μεταβλητές, με μια stepwise-forward διαδικασία, τις οποίες συμβολίζουμε με $\bar{L}$. Μοντελοποιήθηκε, λοιπόν, η $P(Y = 1 \mid \bar{L})$ ως εξής:

$$\log(\frac{P(Y = 1 \mid \bar{L}; \beta)}{1 - P(Y = 1 \mid \bar{L}; \beta)}) = \beta_0 + \sum_{j=1}^{50} \beta_j L_j$$

Στην συνέχεια για κάθε άτομο από το δείγμα, με χρήση της εκτιμήτριας μέγιστης πιθανοφάνειας $\hat{\beta}$ υπολογίστηκε ο προβλεπόμενος από το μοντέλο μέσος $b_i = b_i(\bar{L}_i; \hat{\beta})$ Ο διορθωμένος επιπολασμός της συνεύρεσης επί πληρωμή υπολογίστηκε ως εξής:

$$\hat{\mu}_{OR} = \frac{\sum_{i=1}^{N} W_{i,s} b_i}{\sum_{i=1}^{N} W_{i,s}}$$

Διπλά ανθεκτική εκτιμήτρια Οι εκτιμήσεις με τις δύο παραπάνω μεθόδους θα είναι αμερόληπτες υπό την προϋπόθεση ότι τα μοντέλα είναι σωστά ορισμένα. Καθώς όμως δεν μπορούμε να είμαστε σίγουροι σχετικά με αυτή την υπόθεση εφαρμόζεται η μέθοδος της διπλά ανθεκτικής εκτίμησης (Bang & Robins 2005).

Ορίζεται μια ψευδό-μεταβλητή την οποία χρησιμοποιούμε ως εξαρτημένη ως εξής:

$$\hat{Y}_{i,DR} = \frac{R_i}{\pi_i(\bar{M}; \hat{a})} Y_i - \frac{R_i}{\pi_i(\bar{M}; \hat{a})} b_i + b_i$$

Ο εκτιμητής ορίζεται ως εξής:

$$\hat{\mu_{DR}} = \frac{\sum_{i=1}^{N} W_{i,s} \hat{Y}_{i,DR}}{\sum_{i=1}^{N} W_{i,s}}$$

Κεφάλαιο 3

Μη ανταπόκριση σε επίπεδο ερώτησης

3.1 Εισαγωγή στην αντιμετώπιση της μη ανταπόκρισης σε επίπεδο ερώτησης

3.1.1 Απλές μέθοδοι

Complete Case Analysis Όπως και στην περίπτωση της μη ανταπόκρισης ατόμου, η πρώτη εύκολη λύση, είναι να αγνοήσουμε την ύπαρξη ελλειπουσών τιμών και να κάνουμε την ανάλυση με χρήση μόνο των πλήρων παρατηρήσεων. Διαγράφουμε λοιπόν τις παρατηρήσεις εκείνες που περιέχουν ελλείπουσες τιμές. Η μέθοδος αυτή έχει αρκετά προβλήματα. Είναι έγκυρη μόνο υπό την υπόθεση MCAR. Επομένως, μόνο όταν οι ελλείπουσες τιμές συμβαίνουν κατά εντελώς τυχαίο τρόπο, η πιθανότητα ελλειπουσών τιμών δηλαδή δεν εξαρτάται ούτε από άλλες παρατηρήρηθείσες ούτε από μη παρατηρηθείσες τιμές, η ανάλυση αυτή είναι έγκυρη και δίνει αμερόληπτες εκτιμήσεις. Σε διαφορετική περίπτωση οι εκτιμήσεις θα περιέχουν σφάλμα.

Αντικατάσταση των ελλείπουσων τιμών με τον μέσο Μια σχετικά απλή ιδέα για την αντιμετώπιση των ελλειπουσών τιμών είναι η αντικατάσταση τους με κάποιες αληθοφανείς τιμές, και στην συνέχεια ανάλυση ενός πλήρους αρχείου δεδομένων με τους κλασικούς τρόπους. Μπορούμε λοιπόν να αντικαταστήσουμε τις ελλείπουσες τιμές μιας μεταβλητής με την μέση τιμή της. Αν και φαίνεται λογική η αντιμετώπιση αυτή έχει πολλά μειονεκτήματα. Καταρχάς, δεν είναι εφαρμόσιμη όταν πρόκειται για κατηγορικές μεταβλητές. Επίσης, η κατανομή των μεταβλητών μετά την αντικατάσταση διαστρεβλώνεται καθώς υπάρχει μεγάλη συγκέντρωση τιμών στη μέση τιμή, με αποτέλεσμα την λανθασμένη εκτίμηση, συνήθως υποεκτίμηση, της διασποράς.

Μια παραλλαγή της μεθόδου είναι η αντικατάσταση με τον προβλεπόμενο μέσο από ένα μοντέλο παλινδρόμησης. Κατά την εφαρμογή αυτής της μεθόδου χρησιμοποιούμε τις πλήρεις παρατηρήσεις για να εφαρμόσουμε παλινδρόμηση της μεταβλητής που μας ενδιαφέρει να αντικαταστήσουμε. Άλλες μεταβλητές οι οποίες είναι πλήρεις, μπορούν να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές. Στην συνέχεια αντικαθιστούμε τις παρατηρήσεις που λείπουν με τον προβλεπόμενο από το μοντέλο μέσο. Με αυτόν τον τρόπο κάνουμε χρήση της πληροφορίας για την από κοινού κατανομή των μεταβλητών. Αν και είναι καλύτερη μέθοδος συγκριτικά με την απλή αντικατάσταση με τον μέσο, το ίδιο πρόβλημα παραμένει. Οι εκτιμήσεις για την διασπορά θα είναι εσφαλμένες καθώς οι αντικαταστάσεις δεν αντανακλούν την πραγματική μεταβλητότητα.

Ανεξάρτητα από τον τρόπο με τον οποίο γίνεται η αντικατάσταση με εκτιμώμενες τιμές, αυτή η μέθοδος έχει πολλούς περιορισμούς. Σύμφωνα με τους Little και Rubin (Rubin 1987), οι περιορισμοί είναι ότι:

- Το μέγεθος του δείγματος υπερεκτιμάται.
- Η διασπορά υποτιμάται.
- Οι συσχετίσεις μεταξύ μεταβλητών συνήθως υποεκτιμώνται.
- Η κατανομή των νέων τιμών είναι μια ανακριβής αντιπροσώπευση των τιμών του πληθυσμού, επειδή η μορφή της κατανομής διαστρεβλώνεται με την προσθήκη των τιμών οι οποίες ταυτίζονται με το μέσο όρο.

Η μεροληψία που εισάγεται στη διασπορά του πληθυσμού, στον συντελεστή συσχέτισης δύο μεταβλητών και στην κατανομή των μεταβλητών, εξαρτάται από το ποσοστό των δεδομένων που λείπουν. Οι Little και Rubin συνιστούν να μην χρησιμοποιείται ποτέ η μέθοδος.

Hot deck imputation Η μέθοδος hot deck imputation (hot deck αντικατάσταση) περιλαμβάνει την αντικατάσταση των ελλειπουσών τιμών μιας ή και περισσότερων μεταβλητών ενός μη ανταποκριθέντα (δέκτης), με παρατηρηθείσες τιμές από κάποιον ανταποκριθέντα (δότης), οι οποίοι είναι παρόμοιοι ως προς κάποια παρατηρηθέντα χαρακτηριστικά (Rebecca R. Andridge 2010). Ουσιαστικά πρόκειται για αντιγραφή των τιμών οι οποίες έχουν παρατηρηθεί. Η επιλογή του δότη μπορεί να γίνει με διάφορους τρόπους. Κατά την τυχαία hot deck μέθοδο ο δότης επιλέγεται τυχαία μέσα από ένα σύνολο πιθανών δοτών (donor pool) ενώ κατά την ντετερμινιστική hot deck , ο δότης, ο οποίος αναφέρεται και ως κοντινότερος γείτονας, αναγνωρίζεται με βάση κάποιο μετρικό. Η πλέον διαδεδομένη hot deck μέθοδος είναι η Predictive Mean Matching (PPM).

Η Predictive Mean Matching (PPM) μπορεί να χρησιμοποιηθεί ως εναλλακτική της αντικατάστασης με γραμμική παλινδρόμηση όταν η κατανομή της μεταβλητής δεν είναι κανονική. Η μέθοδος είναι μερικώς παραμετρική και χρησιμοποιεί την απλή γραμμική παλινδρόμηση για την πρόβλεψη τιμών. Στην συνέχεια χρησιμοποιεί τις προβλεπόμενες τιμές ως μέτρο απόστασης για την δημιουργία ενός συνόλου πιθανών δοτών (κοντινότεροι γείτονες). Στο τέλος επιλέγεται τυχαία μια τιμή από το σύνολο αυτό. Το μέγεθος του συνόλου των πιθανών δοτών επιλέγεται από τον αναλυτή ανάλογα με τα δεδομένα, και επηρεάζει την συσχέτιση μεταξύ των αντικαταστάσεων. Όσο μικρότερο το μέγεθος αυτό τόσο υψηλότερη η συσχέτιση και τόσο υψηλότερη η διακύμανση των εκτιμήσεων. Εάν αντίθετα ο αριθμός των πιθανών δοτών είναι πολύ μεγάλος μπορεί να καταλήξουμε σε αυξημένη μεροληψία των εκτιμήσεων. Δεν υπάρχει στην βιβλιογραφία σαφής απάντηση σχετικά με τον κατάλληλο αριθμό δοτών (Schenker & Taylor 1996).

Η μέθοδος αυτή έχει ένα σημαντικό πλεονέκτημα, και αντίθετα με τις υπόλοιπες hot deck imputation μεθόδους, χρησιμοποιείται στην πράξη. Εκτός του ότι είναι η καλύτερη λύση για συνεχείς μεταβλητές που δεν είναι κανονικές, λόγω του ότι αντιγράφει τιμές από άλλες παρατηρήσεις, οι αντικαταστάσεις δεν θα ξεφύγουν ποτέ από το παρατηρούμενο εύρος τιμών.

3.1.2 Μεθοδοι βασισμένες στην πιθανοφάνεια

ΕΜ Αλγόριθμος Ο αλγόριθμος ΕΜ (Expetation-Maximization Εκτίμηση-Μεγιστοποίηση) αποτελεί μία μέθοδο για την επεξεργασία των ελλείπουσων τιμών. Ο όρος ΕΜ καθιερώθηκε από τους Dempster, Laird και Rubin(Dempster et al. 1977). Η διαδικασία ΕΜ περιγράφεται παρακάτω (Καλπινέλλη Ευαγγελία 2004):
(1) Αντικατάσταση των ελλειπουσών τιμών με τις εκτιμημένες τιμές,
(2) εκτίμηση των παραμέτρων,

(3) επανεκτίμηση των ελλειπουσών τιμών υποθέτοντας ότι οι νέες εκτιμήσεις των παραμέτρων είναι σωστές,
(4) επανεκτίμηση των παραμέτρων,
και ούτω καθεξής, επαναλαμβάνοντας την προαναφερθείσα διαδικασία μέχρι να επιτευχθεί σύγκλιση. Κάθε επανάληψη του αλγορίθμου EM αποτελείται από δύο βήματα: ένα βήμα E (Expectation-Εκτίμηση) που ακολουθείται από ένα βήμα M (Maximization-Μεγιστοποίηση). Στο βήμα E, η αναμενόμενη τιμή του λογαρίθμου της πιθανοφάνειας του πλήρους σετ δεδομένων προκύπτει, λαμβάνοντας υπόψη τα παρατηρηθέντα στοιχεία και τις προηγουμένως εκτιμημένες παραμέτρους. Στο βήμα M, η δεσμευμένη αναμενόμενη τιμή του λογαρίθμου της πιθανοφάνειας του πλήρους σετ δεδομένων μεγιστοποιείται. Η τιμή αυτή αυξάνεται έως ότου επιτυγχάνεται ένα στάσιμο σημείο (Dempster et al. 1977). Ο αλγόριθμος συνεχίζεται δηλαδή, έως ότου η παρατηρηθείσα πιθανοφάνεια που παράγεται σε δύο διαδοχικές επαναλήψεις είναι σχεδόν ίδια.

Μέγιστη Πιθανοφάνεια Στην περίπτωση που τα δεδομένα δεν περιέχουν ελλείπουσες τιμές η βασική διαδικασία για την εκτίμηση ενός διανύσματος β περιλαμβάνει την επίλυση της εξίσωσης

$$U(\hat{\beta}; \mathbf{Y}) = 0 \tag{3.1.1}$$

όπου το $\boldsymbol{\Upsilon}$ περιλαμβάνει την μεταβλητή απόκρισης αλλά και τις επεξηγηματικές μεταβλητές. Η μέθοδος αυτή μας δίνει εκτιμήσεις και για τον πίνακα συνδιακύμανσης του $\hat{\beta}$.

Στην περίπτωση που τα δεδομένα περιέχουν ελλείπουσες τιμές η διαδικασία είναι η εξής. Ορίζουμε ως $\mathbf{R}$ το διάνυσμα δείκτη του οποίου τα στοιχεία παίρνουν την τιμή 0 εάν το αντίστοιχο στοιχείο του $\boldsymbol{\Upsilon}$ λείπει και 1 διαφορετικά. Ορίζουμε ως $\mathbf{Y^m}$ το διάνυσμα των στοιχείων του $\mathbf{Y}$ που λείπουν , και ως $\mathbf{Y^o}$ τα παρατηρηθέντα. Ένας συνεπής εκτιμητής από τα μη-πλήρη δεδομένα μπορεί να υπολογιστεί με την επίλυση της εξίσωσης:

$$E_{\mathbf{Y^m}|\mathbf{Y^o},\mathbf{R}}\{U(\hat{\beta}; \mathbf{Y^m}, \mathbf{Y^o})\} = 0 \tag{3.1.2}$$

Από την επίλυση της παραπάνω εξίσωσης μπορεί να υπολογιστεί συνεπής εκτιμήτρια του πίνακα συνδιακύμανσης.

Η μέθοδος αυτή έχει συχνά υπολογιστικές δυσκολίες που την καθιστούν δύσκολη στην εφαρμογή. Για τέτοιες περιπτώσεις άλλες μέθοδοι απαιτούνται.

Κάποιες από τις παραπάνω μεθόδους (οι απλές), αν και δημοφιλείς στην πράξη, κάνουν την εσφαλμένη υπόθεση ότι γνωρίζουμε την τιμή που λείπει. Πιο σωστή προσέγγιση θα ήταν μια πιό στοχαστική προσέγγιση, δοθέντων δηλαδή των παρατηρηθέντων δεδομένων τι θα μπορούσαμε να συμπεράνουμε για την κατανομή των ελλειπουσών· Τέτοιου είδους προσέγγιση είναι οι Μπεϋζιανές Πολλαπλές Αντικαταστάσεις (ή απλά Πολλαπλές Αντικαταστάσεις). Λόγω του καλά τεκμηριωμένου μαθηματικού υπόβαθρου, και την ευκολία στην εφαρμογή της μεθόδου (πλέον με την ανάπτυξη κατάλληλου λογισμικού) είναι η πλέον διαδεδομένη μέθοδος για την αντιμετώπιση ελλειπουσών τιμών. Οι Μπεϋζιανές Πολλαπλές Αντικαταστάσεις είναι το αντικείμενο της επόμενης ενότητας.

3.2 Πολλαπλές αντικαταστάσεις

Multiple imputations

3.2.1 Μπεϋζιανό υπόβαθρο

Στην Μπεϋζιανή στατιστική η συμπερασματολογία γίνεται πάντα μέσω της εκ των υστέρων συνάρτησης κατανομής των παραμέτρων. Με εφαρμογή του θεωρήματος Bayes στις παραμέτρους

ενός όχι πλήρους αρχείου δεδομένων έχουμε:

$$\pi(\theta \mid \mathbf{X}_{obs}) = \frac{\pi(\theta) f(\mathbf{X}_{obs} \mid \theta)}{f(\mathbf{X}_{obs})} \tag{3.2.1}$$

όπου το πρώτο μέλος καλείται εκ των υστέρων κατανομή των παραμέτρων δοθέντων των παρατηρηθέντων δεδομένων ενώ με $\pi(\theta)$ συμβολίζεται η εκ των προτέρων κατανομή των παραμέτρων. Η κατανομή των παρατηρηθέντων δεδομένων $f(\mathbf{X}_{obs})$ μπορεί να θεωρηθεί ως σταθερά κανονικοποίησης επομένως η παραπάνω εξίσωση μπορεί να γραφτεί ως εξής:

$$\pi(\theta \mid \mathbf{X}_{obs}) \propto \pi(\theta) f(\mathbf{X}_{obs} \mid \theta) \tag{3.2.2}$$

Επιπλέον η Μπεϋζιανή συμπερασματολογία, επιτρέπει την πρόβλεψη μελλοντικών παρατηρήσεων με χρήση των παρατηρηθέντων μέσω της εκ των υστέρων συνάρτησης πρόβλεψης. Οι ελλείπουσες τιμές μπορούν να θεωρηθούν ως μελλοντικές τιμές, επομένως είναι δυνατό να υπολογιστεί η συνάρτηση πρόβλεψης ως εξής:

$$f(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}) = \int f(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}, \theta) \pi(\theta \mid \mathbf{X}_{obs}) d\theta \tag{3.2.3}$$

3.2.2 Πολλαπλές αντικαταστάσεις

Στην περίπτωση που όλες ή ένα σύνολο μεταβλητών λείπουν για κάποιες παρατηρήσεις, η στάθμιση με την αντίστροφη πιθανότητα είναι μια εύκολα εφαρμόσιμη και αποτελεσματική μέθοδος. Σε περίπτωση όμως που λείπουν μεμονωμένες τιμές στις διάφορες παρατηρήσεις είναι προτιμότερο να εφαρμοστούν μέθοδοι πολλαπλής αντικατάστασης (MI). Η διαδικασία περιλαμβάνει τα εξής βήματα:

1. Αντικατάσταση των ελλειπουσών τιμών με ένα σύνολο αληθοφανών τιμών για την δημιουργία ενός αριθμού M ψευτο-πλήρων σετ δεδομένων.
2. Ανάλυση των διαφορετικών σετ δεδομένων με χρήση κλασικών μεθόδων και εξαγωγή εκτιμήσεων.
3. Σύνθεση των επιμέρους εκτιμήσεων σε μία.

Η συνδυασμένη εκτίμηση προκύπτει ως απλός μέσος όρος των επιμέρους εκτιμήσεων, δηλαδή συμβολίζοντας ως $\hat{\beta}_{MI}$ την συνθετική εκτίμηση, και $\hat{\beta}_k$ την εκτίμηση από το κ-οστό ψευτο-πλήρες σετ δεδομένων έχουμε:

$$\hat{\beta}_{MI} = \frac{1}{M} \sum_{k=1}^{M} \hat{\beta}_k$$

Η εκτίμηση του πίνακα συνδυακύμανσης είναι ελαφρώς πιο περίπλοκη καθώς θεωρούμε ότι υπάρχουν δύο πηγές αβεβαιότητας, μια εντός αντικαταστάσεων (within imputation) και μια μεταξύ των αντικαταστάσεων (between imputation) . Η εντός αντικαταστάσεων διακύμανση ορίζεται ως εξής:

$$V_w = \frac{1}{M} \sum_{k=1}^{M} V_k$$

όπου V_k ο πίνακας συνδυακύμανσης που προκύπτει από την ανάλυση του κ-οστού ψεύτο-πλήρους σετ δεδομένων.

Η μεταξύ των αντικαταστάσεων διακύμανση προκύπτει ως εξής:

$$V_B = \frac{1}{M-1}\sum_{k=1}^{M}(\hat{\beta}_k - \hat{\beta}_{MI})(\hat{\beta}_k - \hat{\beta}_{MI})'$$

και τελικά η εκτίμηση του πίνακα συνδιακύμανσης δίνεται από την σχέση:

$$V_{MI} = V_W + \frac{M+1}{M}V_B$$

Ο τρόπος με τον οποίο θα υλοποιηθεί το βήμα 1 είναι ιδιαίτερα σημαντικός και γίνεται με χρήση της συνάρτησης πρόβλεψης όπως ορίζεται Μπεϋζιανά. Αρχικά ορίζουμε την εκ των υστέρων συνάρτηση πρόβλεψης με χρήση των παρατηρηθέντων τιμών. Στην συνέχεια δειγματοληπτούμε από αυτήν, και αντικαθιστούμε τις ελλείπουσες τιμές με τις τιμές που προκύπτουν από την δειγματοληψία αυτή. Για τον σκοπό αυτό πρέπει να οριστεί ένα μοντέλο αντικατάστασης (imputation model). Παρακάτω θα αναφέρουμε κάποιες μεθόδους για την δημιουργία του μοντέλου αυτού, πρώτα όμως θα δώσουμε έναν χρήσιμο ορισμό.

Η μονοτονική μορφή : Αν στις μεταβλητές $X_1, X_2, ..., X_n$ υπάρχουν ελλείπουσες τιμές και υπάρχει κάποια αναδιάταξη των μεταβλητών $X'_1, X'_2, ..., X'_n$ τέτοια ώστε η έλλειψη τιμής στην X'_k να συνεπάγεται έλλειψη τιμής στις $X'_{k+1}, ..., X'_n$, τότε οι ελλείπουσες τιμές παρουσιάζουν μονοτονική μορφή.

Η ιδιότητα αυτή είναι ιδιαίτερα χρήσιμη στην επιλογή της μεθόδου αντικατάστασης, όμως στην πράξη σπάνια συναντάται.

3.2.3 Ακολουθιακές αντικαταστάσεις

Στην περίπτωση που οι ελλείπουσες τιμές ακολουθούν μονοτονική μορφή, οι αντικαταστάσεις πολλών μεταβλητών, είναι ισοδύναμες με μια ακολουθία δεσμευμένων μονοπαραγοντικών αντικαταστάσεων. Εάν οι μεταβλητές $X_1, X_2, ..., X_p$ είναι ακολουθούν μονοτονική μορφή, τότε η αντικαταστάσεις μπορούν να γίνουν ακολουθιακά ως εξής:

$$X_1^* \sim f_1(X_1 \mid \mathbf{Z})$$
$$X_2^* \sim f_2(X_2 \mid X_1^*, \mathbf{Z})$$
$$...$$
$$X_p^* \sim f_p(X_p \mid X_1^*, X_2^*, ..., X_{p-1}^*, \mathbf{Z})$$

όπου $\mathbf{Z}$ ένα σύνολο ανεξάρτητων μεταβλητών και $f_j(.)$ η κατάλληλη συνάρτηση ανάλογα με το είδος των δεδομένων (κανονική, λογιστική κ.α.).

Πρώτα λοιπόν γίνεται η αντικατάσταση των τιμών της μεταβλητής εκείνης η οποία έχει τις λιγότερες ελλείπουσες τιμές με την βοήθεια κάποιων ανεξάρτητων μεταβλητών. Στην συνέχεια γίνεται η αντικατάσταση των τιμών της μεταβλητής εκείνης η οποία έχει τις αμέσως επόμενες λιγότερες ελλείπουσες τιμές με την βοήθεια κάποιων ανεξάρτητων μεταβλητών και της ολοκληρωμένης προηγούμενης μεταβλητής. Συνεχίζοντας έτσι δημιουργείται ένα ολοκληρωμένο σετ δεδομένων.

Αυτή η μέθοδος είναι ιδιαίτερα ευέλικτη διότι επιτρέπει την σωστή-κατάλληλη μοντελοποίηση της κάθε μεταβλητής. Όμως στην πράξη είναι δύσκολα εφαρμόσιμη μιας και σπάνια συναντάται μονοτονική μορφή στις ελλείπουσες τιμές. Άλλες μέθοδοι μπορεί να χρησιμοποιηθούν στις περιπτώσεις αυτές.

3.2.4 Χρήση πολυμεταβλητής κατανομής

Είδαμε εν συντομία τον τρόπο με τον οποίο είναι δυνατό να γίνουν πολλαπλές αντικαταστάσεις όταν οι ελλείπουσες τιμές παρουσιάζουν μονοτονική μορφή. Η ίδια μέθοδος δεν μπορεί να εφαρμοστεί όταν τα δεδομένα δεν έχουν αυτήν την ιδιότητα. Η ακολουθιακή αντικατάσταση είναι εφικτή διότι οι μεταβλητές μπορούν να διαταχθούν με τρόπο ώστε οι πλήρεις παρατηρήσεις της μεταβλητής στην οποία γίνεται αντικατάσταση να είναι πλήρεις σε όλες τις προηγούμενες μεταβλητές στις οποίες έχει γίνει αντικατάσταση. Έτσι οι εκτιμήσεις των παραμέτρων από τα συμπληρωμένα δεδομένα δεν εξαρτώνται από τις προηγούμενες αντικαταστάσεις. Αυτό δεν συμβαίνει σε μη μονότονα δεδομένα. Όταν υπάρχει τυχαίο μοτίβο ελλειπουσών τιμών, δεν υπάρχει τέτοια διάταξη έτσι οι εκτιμήσεις εξαρτώνται από τις προηγούμενες αντικαταστάσεις.

Ας θεωρήσουμε δυο μεταβλητές X_1 και X_2. Η X_1 είναι πλήρης στην παρατήρηση 1 και λείπει η παρατήρηση 2. Αντίθετα η X_2 λείπει από την 1 και είναι πλήρης στην 2. Εάν κάνουμε αντικατάσταση της πρώτης παρατήρησης της X_2 με χρήση της αντικατάστασης που έχει γίνει στην X_1 , χειριζόμαστε την τιμή της X_1 που έχει προκύψει από αντικατάσταση ως παρατηρηθείσα πραγματική τιμή αγνοώντας την μεταβλητότητα λόγω αντικατάστασης. Για να λάβουμε υπόψιν μας αυτή τη μεταβλητότητα πρέπει να κάνουμε χρήση κάποιου επαναληπτικού αλγορίθμου ώστε οι τελικές εκτιμήσεις να εξαρτώνται μόνο από τα παρατηρηθεντα δεδομένα, και όχι από τις τιμές που έχουν προκύψει από αντικατάσταση.

Δύο προσεγγίσεις υπάρχουν για την επίλυση του προβλήματος αυτού, η από κοινού μοντελοποίηση και η χρήση αλυσιδωτών εξισώσεων.

Με την από κοινού μοντελοποίηση υποθέτουμε μια πολυμεταβλητή κατανομή για όλες τις μεταβλητές και η αντικατάσταση γίνεται με δειγματοληψία από την εκ των υστέρων συνάρτηση πρόβλεψης των ελλειπουσών τιμών δοθέντων των παρατηρηθέντων. Συχνά είναι δύσκολο να γίνει δειγματοληψία απευθείας από την εκ των υστέρων συνάρτηση πρόβλεψης και σε αυτές τις περιπτώσεις γίνεται χρήση κάποιας μεθόδου MCMC .

3.2.5 Αντικαταστάση με χρήση αλισυδωτών εξισώσεων Multiple imputations using Chained Equations(MICE)

Με αυτή τη μέθοδο αντί να υποθέσουμε μια πολυμεταβλητή από κοινού κατανομή, υποθέτουμε μονομεταβλητες κατανομές για κάθε μια μεταβλητή ξεχωριστά, κατάλληλες για τους διαφορετικούς τύπους των μεταβλητών. Οι υπόλοιπες μεταβλητές (στις οποίες μας ενδιαφέρει να κάνουμε αντικατάσταση) χρησιμοποιούνται ως επεξηγηματικές μεταβλητές. Όπως αναφέρθηκε παραπάνω είναι απαραίτητο να χρησιμοποιηθεί κάποιος επαναληπτικός αλγόριθμος ώστε οι εκτιμήσεις να εξαρτώνται μόνο από τα παρατηρηθεντα δεδομένα. Εάν οι μεταβλητές $X_1, X_2, ..., X_p$ περιέχουν ελλείπουσες τιμές και οι μεταβλητές $\mathbf{X}_{obs}$ είναι πλήρεις, το μοντέλο για την X_k μεταβλητή είναι:

$$f(X_k \mid X_1, X_2, ..., X_{k-1}, X_{k+1}, ..., X_p, \mathbf{X}_{obs}), k = 1, ..., p$$

Ένας αλγόριθμος τύπου Gibbs εφαρμόζεται στη συνέχεια σε αυτά τα μονομεταβλητά μοντέλα, ο οποίος σε αντίθεση με τον παραδοσιακό αλγόριθμο Gibbs εφαρμόζεται μόνο στις παρατηρήσεις εκείνες με παρατηρηθείσα απόκριση X_k . Επαναλαμβάνοντας κυκλικά σε όλα τα μοντέλα στη σειρά, λαμβάνονται τιμές από τις εκ των υστέρων κατανομές, με δεδομένες τις τρέχουσες τιμές για τις άλλες μεταβλητές. Η διαδικασία μπορεί να γραφτεί ως εξής για την t-οστή επανάληψη:

$$X_1^{(t+1)} \sim g_1(X_1 \mid X_2^{(t)}, ..., X_p^{(t)}, \mathbf{Z}, \phi_1)$$

$$X_2^{(t+1)} \sim g_2(X_2 \mid X_1^{(t+1)}, X_3^{(t)}, ..., X_p^{(t)}, \mathbf{Z}, \phi_2)$$

...

$$X_p^{(t+1)} \sim g_p(X_p \mid X_1^{(t+1)}, X_2^{(t+1)}, ..., X_{p-1}^{(t+1)}, \mathbf{Z}, \phi_p)$$

όπου ϕ_j οι παράμετροι του αντίστοιχου μοντέλου.

Αρχικές τιμές είναι απαραίτητες για τον πρώτο κύκλο. Συνήθως αυτές λαμβάνονται με απλή τυχαία δειγματοληψία με επανάθεση από τις παρατηρηθείσες τιμές. Στη συνεχεία γίνεται παλινδρόμηση της πρώτης μεταβλητής με ελλείπουσες τιμές X_1, σε όλες τις άλλες μεταβλητές $X_2, ..., X_p$, μόνο στις παρατηρήσεις εκείνες που έχουν παρατηρηθείσα τιμή για το X_1. Οι ελλείπουσες τιμές της X_1 αντικαθίστανται με προσομοιωμένες τιμές από την αντίστοιχη εκ των υστέρων κατανομή της X_1. Στη συνεχεία γίνεται παλινδρόμηση της δεύτερης μεταβλητής με ελλείπουσες τιμές X_2, σε όλες τις άλλες μεταβλητές $X_1, Q_3, ..., X_p$, μόνο στις παρατηρήσεις εκείνες που έχουν παρατηρηθείσα τιμή για το X_2 και με χρήση των τιμών της X_1 που έχουν προκύψει από την προηγούμενη αντικατάσταση. Οι ελλείπουσες τιμές της X_2 αντικαθίστανται με προσομοιωμένες τιμές από την αντίστοιχη εκ των υστέρων κατανομή της X_2. Η διαδικασία επαναλαμβάνεται για τις $Q_3, ..., Q_p$, αυτό καλείται ένας κύκλος. Η διαδικασία επαναλαμβάνεται για αρκετούς κύκλους (π.χ. 10 ή 20), ώστε οι τελικές εκτιμήσεις να εξαρτώνται μόνο από τα παρατηρηθεντα δεδομένα, και όχι από τις τιμές που έχουν προκύψει από αντικατάσταση. Μετά από τον τελευταίο κύκλο, οι τρέχουσες τιμές χρησιμοποιούνται για την εξαγωγή ενός πλήρους σετ δεδομένων. Η όλη διαδικασία επαναλαμβάνεται M φορές για την εξαγωγή M πλήρων σετ δεδομένων.

Τα πλεονεκτήματα αυτής της μεθόδου είναι ότι η κάθε μεταβλητή μοντελοποιείται καταλλήλως και είναι σχετικά απλό να περιληφθούν αλληλεπιδράσεις και μη-γραμμικοί όροι στο μοντέλο. Το μειονέκτημα είναι η έλλειψη θεωρητικής τεκμηρίωσης. Υπάρχουν περιπτώσεις στις οποίες οι δεσμευμένες πυκνότητες $g_j(.), j = 1, ..., p$ δεν αντιστοιχούν σε μια από κοινού πολυμεταβλητή κατανομή των $X_1, X_2, ..., X_p$. Σε αυτές τις περιπτώσεις, οι δεσμευμένες κατανομές δεν είναι συμβατές και ο αλγόριθμος δεν θα καταλήξει σε μια στάσιμη κατανομή(Arnold et al. 1999). Το αντίκτυπο αυτής της αδυναμίας της μεθόδου, είναι ακόμα υπό διερεύνηση. Οι van Buuren et al. (2006) μέσω προσομοιώσεων, συμπέραναν ότι, σε απλές περιπτώσεις, το αντίκτυπο του προβλήματος αυτού στις εκτιμήσεις είναι μικρό (Van Buuren et al. 2006).

Η διαδικασία είναι ένας επαναληπτικός αλγόριθμος όμοιος με τον δειγματολήπτη Gibbs (Gelfand & Smith 1990). Ο δειγματολήπτης Gibbs είναι ένας αλγόριθμος ο οποίος χρησιμοποιείται όταν είναι δύσκολο να γίνει ευθεία προσομοίωση από κάποια πολυμεταβλητή κατανομή. Ο Gibbs προσωμοιώνει τιμές από τις δεσμευμένες κατανομές των παραμέτρων, και κάτω από κάποιες συνθήκες ομαλότητας και μετά από μια περίοδο burn-in, καταλήγει να παίρνει δείγμα από την κατανομή στόχο, η οποία ονομάζεται στάσιμη κατανομή. Τότε λέμε ότι έχει επιτευχθεί σύγκλιση του αλγορίθμου. Στο ίδιο πνεύμα, κατά την εφαρμογή της μεθόδου MICE απαιτούνται κάποιες burn-in επαναλήψεις, καθώς και εποπτεία της σύγκλισης του αλγορίθμου. Η σύγκλιση του αλγορίθμου εξετάζεται γραφικά (StataCorp 2013).

Λόγω της ευελιξίας της μεθόδου και παρά την έλλειψη θεωρητικής τεκμηρίωσης, είναι ιδιαίτερα διαδεδομένη. Θα δούμε παρακάτω πως εφαρμόζεται πρακτικά η μέθοδος αυτή.

3.2.6 MICE στην πράξη

Μοντελοποίηση των μετβλητών

Συνεχείς μεταβλητές Για συνεχείς μεταβλητές ένα γραμμικό μοντέλο παλινδρόμησης είναι η συνήθης επιλογή μοντέλου για κανονικά κατανεμημένες μεταβλητές.

$$x \mid \mathbf{z}, \beta \sim N(\beta \mathbf{z}, \sigma^2)$$

Έστω $\hat{\beta}$ οι εκτιμήσεις (διάνυσμα μήκους κ) από την εφαρμογή του μοντέλου στις παρατηρήσεις με παρατηρηθέν x , και n_{obs} το πλήθος των παρατηρήσεων αυτών. Επίσης έστω $\mathbf{V}$ η εκτίμηση του πίνακα συνδυακύμανσης των $\hat{\beta}$, και $\hat{\sigma}$ η ρίζα του μέσου τετραγωνικού σφάλματος.

Πέρνουμε δείγματα των παραμέτρων αντικατάστασης β^*, σ^* από την από κοινού κατανομή των β, σ. Πρώτα το σ^* ως :

$$\sigma^* = \hat{\sigma}\sqrt{(n_{obs} - \kappa)/g}$$

Όπου g είναι μια τυχαία τιμή από μια x^2 κατανομή με $n_{obs} - \kappa$ βαθμούς ελευθερίας. Στην συνέχεια το β^* ως εξής:

$$\beta^* = \hat{\beta} + \frac{\sigma^*}{\hat{\sigma}} u_1 \mathbf{V}^{1/2}$$

όπου u_1 ένα διάνυσμα γραμμή μήκους κ από τυπική κανονική κατανομή και $\mathbf{V}^{1/2}$ η αποσύνθεση Cholevsky του πίνακα $\mathbf{V}$. Οι τιμές για αντικατάσταση x_i^* για κάθε ελλείπουσα τιμή της x_i λαμβάνονται ως εξής:

$$x_i^* = \beta^* z_i + u_{2i}\sigma^*$$

όπου u_{2i} τυχαία τιμή από την τυπική κανονική κατανομή.

Δίτιμες μεταβλητές Το μοντέλο που συνήθως επιλέγεται για την αντικατάσταση δίτιμων μεταβλητών είναι το μοντέλο λογιστικής παλινδρόμησης:

$$logit(Pr(x = 1 \mid \mathbf{z}; \beta)) = \beta' \mathbf{z}$$

Έστω $\hat{\beta}'$ οι εκτιμήσεις από την εφαρμογή του μοντέλου στις παρατηρήσεις με παρατηρηθέν x και $\mathbf{V}'$ η εκτίμηση του πίνακα συνδυακύμανσης. Έστω επίσης $\beta^{*\prime}$ μια τυχαία τιμή από την εκ των υστέρων κατανομή των β , η οποία προσεγγίζεται από μια $MVN(\hat{\beta}, \mathbf{V})$. Για κάθε ελλείπουσα παρατήρηση x_i έστω $p_i^* = [1 + exp(-\beta^{*\prime}\mathbf{z}_i)]^{-1}$, τότε γίνεται αντικατάσταση της παρατήρησης ως εξής :

$$x_i^* = \begin{cases} 1, & \text{αν } u_i < p_i^* \\ 0, & \text{αλλιώς} \end{cases}$$

όπου u_i μια τυχαία τιμή από ομοιόμορφη στο (0,1). Προβλήματα μπορεί να προκύψουν σε περίπτωση τέλειας πρόβλεψης, όταν δηλαδή για κάποια ή κάποιες παρατηρήσεις η προβλεπόμενη από το μοντέλο πιθανότητα είναι ίση με μηδέν ή ένα.

3.2.7 Το μοντέλο Αντικατάστασης

Είναι σημαντικό το μοντέλο αντικατάστασης να διατηρεί όλα τα βασικά χαρακτηριστικά των δεδομένων που έχουν παρατηρηθεί. Αυτό περιλαμβάνει τα ακόλουθα:

1. Χρήση όσο τον δυνατόν περισσότερων ανεξάρτητων μεταβλητών στο μοντέλο , ώστε να αποφευχθούν λανθασμένες υποθέσεις σχετικά με τις σχέσεις μεταξύ των μεταβλητών. Παράβλεψη σημαντικών προγνωστικών μεταβλητών από το μοντέλο αντικατάστασης μπορεί να οδηγήσει σε μεροληπτικές εκτιμήσεις για τις μεταβλητές αυτές στο στάδιο της ανάλυσης. Από την άλλη, αν το μοντέλο αντικατάστασης περιλαμβάνει μη σημαντικούς προγνωστικούς παράγοντες οι εκτιμήσεις θα παραμείνουν έγκυρες θα είναι όμως λιγότερο επαρκείς .

2. Η δομή των δεδομένων θα πρέπει να αντιπροσωπεύεται στο μοντέλο αντικατάστασης. Οι μεταβλητές που είναι σχετικές με τον σχεδιασμό της μελέτης πρέπει να περιληφθούν στο μοντέλο αντικατάστασης. Αν τα δεδομένα προέρχονται από πληθυσμιακή μελέτη, τα δειγματικά βάρη, τα στρώματα και οι συστάδες θα χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές. Αντίστοιχα σε χρονολογικά δεδομένα οι δείκτες για τις επαναλαμβανόμενες μετρήσεις.

3. Προσδιορισμός της σωστής συναρτησιακής μορφής του μοντέλου αντικατάστασης. Για παράδειγμα με χρήση όρων αλληλεπίδρασης.

Το μοντέλο αντικατάστασης θα πρέπει να είναι συμβατό με το μοντέλο που θα χρησιμοποιηθεί στην ανάλυση. Αν μια μεταβλητή χρησιμοποιηθεί στο μοντέλο ανάλυσης θα πρέπει να χρησιμοποιηθεί και στο μοντέλο αντικατάστασης. Αν στην ανάλυση εκτιμάται η συσχέτιση μεταξύ δύο μεταβλητών θα πρέπει και οι δύο να συμπεριληφθούν στο μοντέλο αντικατάστασης. Η μεταβλητή απόκρισης θα πρέπει να περιλαμβάνεται πάντα στο μοντέλο αντικατάστασης. Επιπλέον, εκτός από τις μεταβλητές που μπορεί να περιληφθούν στο μοντέλο ανάλυσης, το μοντέλο αντικατάστασης πρέπει να περιλαμβάνει όλες τις μεταβλητές που μπορεί να περιέχουν πληροφορία σχετική με τα ελλειπή δεδομένα. Αυτό θα κάνει την υπόθεση MAR πιο αληθοφανή και θα βελτιώσει την ποιότητα των τιμών που θα προκύψουν από αντικατάσταση.

Η επιλογή του μοντέλου αντικατάστασης μπορεί να έχει σημαντική επίδραση στα αποτελέσματα ιδίως αν το ποσοστό της ελλείπουσας πληροφορίας είναι υψηλό. Υπάρχουν πολλές διαθέσιμες μέθοδοι για την μοντελοποίηση ανάλογα με τον τύπο των μεταβλητών. Όμως αυτές οι μέθοδοι δεν μπορούν να καλύψουν όλες τις πιθανές κατανομές των δεδομένων που μπορεί να προκύψουν. Συχνά οι μεταβλητές του μοντέλου αντικατάστασης χρειάζεται να μετασχηματιστούν καταλλήλως. Λογαριθμικός μετασχηματισμός (ή γενικότερα Box-Cox μετασχηματισμός) μπορεί να χρησιμοποιηθεί για λοξές συνεχείς κατανομές ώστε αυτές να μοντελοποιηθούν ως κανονικές. Οι μεταβλητές μπορούν να μετασχηματιστούν πίσω στην αρχική κλίμακα μετά την εφαρμογή του μοντέλου. Μετασχηματισμοί είναι χρήσιμοι επίσης όταν θέλουμε να διασφαλίσουμε ότι οι τιμές που θα προκύψουν μετά από αντικατάσταση είναι μεταξύ του 0 και 1. Αυτό είναι εφικτό με χρήση logit μετασχηματισμού. Πρέπει να θυμόμαστε ότι η μοντελοποίηση γίνεται στην δεσμευμένη κατανομή της μεταβλητής επομένως ο τυχόν μετασχηματισμός θα πρέπει να είναι κατάλληλος για αυτήν.

3.2.8 Σχεδιασμός της μελέτης και MI

Η θεωρία των MI για ελλειπή δεδομένα απαιτεί ότι οι αντικαταστάσεις γίνονται λαμβάνοντας υπόψιν τον σχεδιασμό της μελέτης. Όμως συνήθως γίνεται υπόθεση απλής τυχαίας δειγματοληψίας πρακτική που μπορεί να οδηγήσει σε εσφαλμένες εκτιμήσεις ακόμα και όταν οι ανάλυση των πολλαπλών σετ δεδομένων που θα προκύψουν γίνει με βάση τον σχεδιασμό. Μέσω προσομοιώσεων (Reiter et al. 2006) έχει δειχθεί ότι:
α) το σφάλμα που προκύπτει μπορεί να είναι σημαντικό ιδιαίτερα όταν η υπό μελέτη μεταβλητή σχετίζεται με τα χαρακτηριστικά του σχεδιασμού και
β) το σφάλμα μπορεί να μειωθεί όταν ελέγξουμε για τον σχεδιασμό στο μοντέλο αντικατάστασης. Η ίδια μελέτη έδειξε επίσης ότι λαμβάνοντας υπόψιν και άσχετα χαρακτηριστικά του σχεδιασμού στο μοντέλο αντικατάστασης, οι εκτιμήσεις είναι συντηρητικές εάν βέβαια και τα σχετικά χαρακτηριστικά παραμείνουν στο μοντέλο. Η ασφαλέστερη στρατηγική όταν έχουμε να χειριστούμε ελλειπή δεδομένα τα οποία προέρχονται από σύνθετες σχεδιαστικά μελέτες είναι να συμπεριλάβουμε τα χαρακτηριστικά του σχεδιασμού, εκτός από το στάδιο της ανάλυσης, και κατά το στάδιο της αντικατάστασης.

Στις πληθυσμιακές μελέτες ο σύνθετος σχεδιασμός ανταναικλάται στα στρώματα και στις συστάδες στις οποίες έχει γίνει η δειγματοληψία, καθώς επίσης και στα βάρη τα οποία έχουν δημιουργηθεί για κάθε παρατήρηση με βάση τον σχεδιασμό. Αυτές οι μεταβλητές θα πρέπει να εισαχθούν στο μοντέλο αντικατάστασης. Ένας απλός τρόπος είναι να εισαχθούν οι συστάδες και τα στρώματα είναι με χρήση ψευδομεταβλητών ως ανεξάρτητες μεταβλητές στο μοντέλο αντικατάστασης. Εναλλακτικά τα βάρη μπορούν να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές είτε ως συνεχείς, είτε ως κατηγορικές αφού πρώτα χωριστούν σε κατηγορίες σχετικές με την κατανομή τους π.χ. σε τεταρτημόρια. Σε κάθε περίπτωση είναι σημαντικό να ελεγχθούν τυχόν αλληλεπιδράσεις μεταξύ των μεταβλητών της μελέτης και των χαρακτηριστικών του σχεδιασμού. Υπάρχουν όμως

και διαφορετικές πιο περίπλοκες προσεγγίσεις. Μπορεί να χρησιμοποιηθεί μοντέλο αντικατάστασης σταθμισμένο με τα βάρη της μελέτης, είτε, ιεραρχικό μοντέλο με τις συστάδες ως τυχαίες επιδράσεις και τα στρώματα ως σταθερές επιδράσεις (Reiter et al. 2006).

Πιο συγκεκριμένα σύμφωνα με τον Carpenter (Carpenter 2011):

- Η χρήση σταθμισμένου μοντέλου αντικατάστασης αντιμετωπίζει τυχόν σφάλματα στις αντικαταστάσεις, αλλά η σταθμισμένη πιθανοφάνεια δεν αντιστοιχεί σε πιθανοθεωρητικό μοντέλο και η Μπεϋζιανή εκ των υστέρων κατανομή δεν αντιπροσωπεύει την αβεβαιότητα των αντικαταστάσεων.

- Η χρήση των βαρών ως ανεξάρτητες μεταβλητές πρέπει να γίνει με κατάλληλο τρόπο ώστε οι αντικαταστάσεις να μην περιέχουν σφάλμα. Επιπλέον έχει δειχθεί ότι εάν τα βάρη εισαχθούν ως ανεξάρτητες μεταβλητές στο μοντέλο αντικατάστασης και επιπλέον συμπεριληφθούν όροι αλληλεπίδρασης των βαρών με όλες της υπόλοιπες μεταβλητές, τότε οι αντικαταστάσεις για τις ελλείπουσες αποκρίσεις του μοντέλου ανάλυσης είναι ασυμπτωτικά έγκυρες (Kim et al. 2006, Seaman et al. 2012). Παρόλα αυτά οι αντικαταστάσεις για ελλείπουσες τιμές των εξαρτημένων μεταβλητών του μοντέλου ανάλυσης μπορεί να καταλήξουν σε υπέρ εκτίμηση των διακυμάνσεων, αν και μελέτες προσομοίωσης έχουν δείξει ότι αυτή είναι μικρή. Και ένας απλός γραμμικός όρος των βαρών φαίνεται να αρκεί.

3.2.9 Διαγνωστικά

Σύγκλιση του Αλγορίθμου

Οι Μπεϋζιανές πολλαπλές αντικαταστάσεις ουσιαστικά ανάγονται στο πρόβλημα της προσομοίωσης τιμών από την από κοινού εκ των υστέρων συνάρτηση πρόβλεψης. Συνήθως δεν είναι δυνατό να πάρουμε δείγμα με ευθεία προσομοίωση, γι αυτό το λόγο συνήθως χρησιμοποιείται κάποιος αλγόριθμος MCMC. Αυτοί οι αλγόριθμοι είναι επαναληπτικοί και ακολουθιακοί και μετά από κάποιες επαναλήψεις καταλήγουν σε δείγμα από την στάσιμη κατανομή στόχο. Αυτή η κατανομή, στην περίπτωση των αντικαταστάσεων με χρήση αλυσιδωτών εξισώσεων, δεν υπάρχει εγγύηση ότι ορίζεται. Επομένως είναι σημαντικό να γίνει έλεγχος για το κατά πόσον έχει επιτευχθεί η σύγκλιση του αλγορίθμου. Αυτός ο έλεγχος γίνεται γραφικά με χρήση traceplots, και στόχος είναι να μην παρατηρούνται τάσεις στο γράφημα, όπως συνεχή μείωση ή αύξηση. Ένας δεύτερος έλεγχος είναι η γραφική εποπτεία διαφορετικών αλυσίδων και η σύγκριση μεταξύ τους. Οι διαφορετικές αλυσίδες δεν πρέπει να παρουσιάζουν πολύ διαφορετική συμπεριφορά. Αυτό θα ήταν ένδειξη αστάθειας του αλγορίθμου.

Σε κάθε περίπτωση όταν προκύπτουν τέτοια προβλήματα θα πρέπει να γίνει επανεξέταση και τροποποίηση του μοντέλου αντικατάστασης. Σύμφωνα με δημοσιευμένη μελέτη (Bouhlila & Sellaouti 2013) τέτοιου είδους προβλήματα προκύπτουν όταν γίνεται αντικατάσταση σε μεταβλητές που υποθέτουν περίπλοκα μοντέλα όπως πολυωνυμικά. Τα πολυωνυμικά μοντέλα μπορεί να δημιουργήσουν κελιά μικρού μεγέθους. Προτείνεται να επανεξεταστούν οι συσχετίσεις να να περιληφθούν μόνο οι ισχυροί παράγοντες.

Το πρόβλημα της τέλειας πρόβλεψης

Το πρόβλημα αυτό προκύπτει κατά την αντικατάσταση κατηγορικών μεταβλητών (λογιστική παλινδρόμηση κ.α.) όταν για κάποια ή κάποιες παρατηρήσεις η προβλεπόμενη από το μοντέλο πιθανότητα είναι ίση με μηδέν ή ένα και προκαλεί αστάθεια των εκτιμήσεων. Μια λύση στο πρόβλημα είναι να αγνοήσουμε τις μεταβλητές που είναι υπεύθυνες για την τέλεια πρόβλεψη. Έτσι ερχόμαστε σε αντίθεση με την λογική των πολλαπλών αντικαταστάσεων. Μια δεύτερη επιλογή είναι να χειριστούμε αυτό το πρόβλημα με μια τεχνική προσαύξησης των δεδομένων (White et al. 2010) κατά

την οποία επιπλέον παρατηρήσεις με μικρά βάρη προστίθενται στα δεδομένα ώστε καμία πρόβλεψη να μην είναι τέλεια.

Μέρος II

Ανάλυση Δεδομένων

Κεφάλαιο 4

Προετοιμασία των Δεδομένων για ανάλυση

4.1 Δημιουργία βαρών

4.1.1 Τα αρχικά βάρη-δειγματοληπτικά βάρη

Σε προηγούμενα κεφάλαια έχει τονιστεί η ανάγκη της στατιστικής ανάλυσης με βάση τον σχεδιασμό της μελέτης. Για να είναι εφικτή τέτοιου είδους ανάλυση είναι απαραίτητο κάθε παρατήρηση να συνοδεύεται από το αντίστοιχο βάρος. Το βάρος έχει μια πρακτική ερμηνεία. Εκφράζει τον αριθμό των ατόμων του πληθυσμού που αντιπροσωπεύει η κάθε παρατήρηση του δείγματος. Έτσι, εάν ένα άτομο στο δείγμα, έχει βάρος ίσο με 1000, μπορούμε να πούμε ότι υπάρχουν 1000 άτομα στον πληθυσμό με τα χαρακτηριστικά του ατόμου στο δείγμα. Τα βάρη σχεδιάζονται με τρόπο ώστε το άθροισμα τους να είναι ίσο με το μέγεθος του πληθυσμού αναφοράς. Με αυτόν τον τρόπο, μπορούμε να πούμε ότι το σταθμισμένο δείγμα είναι αντιπροσωπευτικό του πληθυσμού. Η δημιουργία των βαρών περιλαμβάνει αρκετά στάδια, ανάλογα με τις διορθώσεις τις οποίες ο ερευνητής επιθυμεί να κάνει. Το πρώτο απαραίτητο βήμα είναι ο υπολογισμός των αρχικών-δειγματοληπτικών βαρών (base weights) τα οποία ορίζονται ως το αντίστροφο της πιθανότητας επιλογής στο δείγμα.

Αν η πιθανότητα επιλογής στο δείγμα είναι π_i για την $i-$οστή παρατήρηση τότε το αντίστοιχο βάρος για την παρατήρηση αυτή είναι απλώς $w_{Bi} = \frac{1}{\pi_i}$. Ανάλογα με τον σχεδιασμό της μελέτης τα αρχικά βάρη μπορεί να είναι πολύ εύκολα στον υπολογισμό καθώς ο υπολογισμός της πιθανότητας επιλογής είναι απλός. Εάν το δείγμα έχει επιλεγεί με ίσες πιθανότητες τότε $\pi_i = n/N$ όπου n το μέγεθος του δείγματος και N το μέγεθος το πληθυσμού από τον οποίο έχει ληφθεί το δείγμα.

Στην περίπτωση της στρωματοποιημένης δειγματοληψίας με ίση πιθανότητα επιλογής ανά στρώμα τότε η πιθανότητα επιλογή της $i-$οστής παρατήρησης στο h στρώμα υπολογίζεται ως εξής:

$$\pi_{Bhi} = n_h/N_h$$

όπου n_h το μέγεθος του δείγματος και N_h το συνολικό μέγεθος του h στρώματος. Το βάρος λοιπόν το οποίο θα λάβει η $i-$οστή παρατήρηση στο h στώμα είναι $w_{Bhi} = N_h/n_h$.

Λίγο πιο περίπλοκος είναι ο υπολογισμός όταν η δειγματοληψία είναι πολυσταδιακή. Αν για παράδειγμα έχουμε τρία στάδια δειγματοληψίας με πρωταρχική δειγματοληπτική μονάδα i, δευτερογενή δειγματοληπτική μονάδα j και τριτογενή δειγματοληπτική μονάδα k τότε η πιθανότητα επιλογής της k μονάδας είναι το γινόμενο των πιθανοτήτων επιλογής σε κάθε στάδιο της δειγματοληψίας. Είναι δηλαδή:

$$\begin{aligned} P(unit\, i,j,k\, selected) &= P(primary\, i\, seected) \\ &\cdot P(secondary\, j\, selected \mid primary\, i\, selected) \\ &\cdot P(tertiary\, k\, selected \mid secondary\, j\, in\, primary\, i\, selected) \end{aligned}$$

Εάν όλες οι δειγματοληπτικές μονάδες σε κάθε στάδιο επιλέγονται με ίση πιθανότητα με επανατοποθέτηση τότε:

$$P(unit\, i,j,k\, selected) = \frac{m}{M} \cdot \frac{m_i}{M_i} \cdot \frac{m_{ij}}{M_{ij}}$$

Το αντίστροφο της παραπάνω ποσότητας είναι το βάρος της κάθε παρατήρησης (Paul S.Levy 2008).

Παράδειγμα

Θα περιγράψουμε τον υπολογισμό των βαρών σε μία στρωματοποιημένη δισταδιακή δειγματοληψία. Η μελέτη National Survey of Child and Adolescent Well-Beign (NSCAW) αφορούσε κακοποιημένα και παραμελημένα παιδιά στις ΗΠΑ. (Dowd et al. 2002) Ο πληθυσμός αποτελείται από όλα τα παιδιά των ΗΠΑ τα οποία υπόκεινται σε κακοποίηση. Για την δειγματοληψία χρησιμοποιήθηκαν κατάλογοι παιδιών για τα οποία υπήρχε ισχυρισμός και διερεύνηση για κακοποίηση ή παραμέληση τον τελευταίο μήνα.

Το δειγματοληπτικό σχέδιο ήταν στρωματοποιημένη δισταδιακή δειγματοληψία. Οι πρωταρχικές δειγματοληπτικές μονάδες ήταν οι περιφέρειες και οι δευτερεύουσες τα παιδιά μέσα στις περιφέρειες τα οποία εκπλήρωναν τα κριτήρια επιλογής. Για δοθέν στρώμα h η πιθανότητα το j παιδί της i περιφέρειας να επιλεγεί στο δείγμα συμβολίζεται με π_{hi} ενω η πιθανότητα να επιλεγεί το παιδί δεδομένου ότι η περιφέρεια έχει επιλεγεί συμβολίζεται με π_{hij}. Το αντίστροφο της πιθανότητας επιλογής μας δίνει το αρχικό βάρος της κάθε παρατήρησης:

$$w_{Bhij} = \frac{1}{\pi_{hi}} \cdot \frac{1}{\pi_{hij}} \tag{4.1.1}$$

Το άθροισμα των παραπάνω βαρών είναι εκτίμηση του συνολικού αριθμού των επιλέξιμων παιδιών στο συγκεκριμένο στρώμα. Το άθροισμα των βαρών στο σύνολο των στρωμάτων αποτελεί εκτίμηση του συνόλου των επιλέξιμων παιδιών στον πληθυσμό. Οι περιφέρειες επιλέχθηκαν χωρίς επανατοποθέτηση με πιθανότητα επιλογής ανάλογη του μεγέθους τους. Έτσι η πιθανότητα επιλογής για την περιφέρεια (h,i) είναι ίση με $m_h M_{hi}/N_h$ όπου:
m_h = αριθμός των περιφερειών που επιλέχθηκαν από το h στρώμα
M_{hi} = ο αριθμός των επιλέξιμων παιδιών της περιφέρειας (h ,i)
N_h = ο συνολικός αριθμός των επιλέξιμων παιδιών στο h στρώμα.
Ομοίως η πιθανότητα επιλογής του (h,i,j) παιδιού ισούται με m_{hi}/M_{hi} όπου:
m_{hi} = ο αριθμός των παιδιών που επιλέχθηκαν στην i περιφέρεια
M_{hi} = ο αριθμός των παιδιών στην i περιφέρεια

Τελικά, με τους παραπάνω συμβολισμούς τα αρχικά βάρη υπολογίζονται ως εξής:

$$w_{Bhi} = \frac{N_h}{m_h M_{hi}} \cdot \frac{M_{hi}}{mhi} = \frac{N_h}{m_h} \cdot \frac{1}{m_{hi}} \tag{4.1.2}$$

4.1.2 Προσαρμογή των αρχικών βαρών Ποσοστά ανταπόκρισης

Όλες οι μελέτες πληθυσμού υποφέρουν από μη ανταπόκριση. Είναι σημαντική η τροποποίηση των δειγματοληπτικών βαρών ώστε να μειωθεί η επίδραση της μη ανταπόκρισης στις εκτιμήσεις μας. Αυτή η διαδικασία απαιτεί γνώση από την πλευρά των ερευνητών χαρακτηριστικών των ατόμων που αρνήθηκαν συμμετοχή στην μελέτη. Όσα περισσότερα τέτοια χαρακτηριστικά είναι γνωστά τόσο πιο λεπτομερειακή διόρθωση μπορεί να γίνει και τόσο πιο μεγάλο μέρος του σφάλματος μπορεί να εξαλειφθεί. Δυστυχώς όμως η πληροφορία αυτή είναι περιορισμένη. Η μόνο πληροφορία που είναι συνήθως γνωστή είναι η περιοχή από την οποία τα άτομα αυτά προέρχονται. Αναμένουμε λοιπόν η διόρθωση που θα γίνει να αφορά παράγοντες οι οποίοι σχετίζονται με την περιοχή όπως κοινωνικοί - οικονομικοί παράγοντες, διατροφή κ.α.

Τα ποσοστά ανταπόκρισης υπολογίζονται ανά δειγματοληπτικό σημείο. Το δειγματοληπτικό σημείο αποτελείται από 2 ή 3 γειτονικά οικοδομικά τετράγωνα ανάλογα με την περιοχή. Ο υπολογισμός είναι απλός. Το ποσοστό ανταπόκρισης, στο δειγματοληπτικό σημείο i, είναι ίσο με το σύνολο των ατόμων στό δείγμα από το συγκεκριμένο δειγματοληπτικό σημείο προς το σύνολο των επιλέξιμων ατόμων που προσεγγίστηκαν (επισκέψεις).

$$\text{Ποσοστό ανταπόκρισης}_i = \frac{\text{Άτομα στο δείγμα}_i}{\text{Επισκέψεις}_i}$$

Από τον παρονομαστή εξαιρούνται όσες επισκέψεις έγιναν στις οποίες διαπιστώθηκε ότι η κατοικία είναι μη επιλέξιμη. Στην EMENO και σύμφωνα με τις οδηγίες για πληθυσμιακές μελέτες μορφή HES (Tolonen et al. 2008) ως μη επιλέξιμη χαρακτηρίζεται μια κατοικία για κάποιον από τους παρακάτω λόγους:
Δευτερεύουσα κατοικία
Ακατοίκητο
Επαγγελματική χρήση
Απουσιάζουν για περισσότερο από 2 εβδομάδες
Απουσίαζαν σε 3 επισκέψεις, μία εκ των οποίων έγινε πρωινές ώρες, μία απογευματινές ώρες από Δευτέρα έως Παρασκευή και μία Σαββατοκύριακο
Απροσδιόριστο.

Αντίθετα ως άρνηση χαρακτηρίζεται μια επίσκεψη για κάποιον από τους παρακάτω λόγους:
Πρόβλημα γλώσσας
Αρνήθηκε: Δεν αιτιολόγησε
Αρνήθηκε: Λόγω έλλειψης χρόνου
Αρνήθηκε: Θέμα προσωπικής αρχής (Συνειδητά δε συμμετέχει σε έρευνες)
Αρνήθηκε: Λόγω ασθενείας
Αρνήθηκε: Αισθάνεται υγιής (έτσι δε νομίζει ότι έχει λόγο να συμμετάσχει)
Αρνήθηκε: Το θέμα της έρευνας (δεν ενδιαφέρεται για θέματα υγείας ή τα θεωρεί πολύ προσωπικά
Αρνήθηκε για άλλο λόγο
Δεν παρουσιάστηκε στη συνάντηση

Με πολλαπλασιασμό του δειγματοληπτικού βάρους με το αντίστροφο του ποσοστού ανταπόκρισης γίνεται η διόρθωση για μη ανταπόκριση. Το νέο αυτό βάρος θα χρησιμοποιηθεί στην ανάλυση. Επομένως εάν συμβολίσουμε με RR_i το ποσοστό ανταπόκρισης, με w_{Bi} το δειγματοληπτικό βάρος και w_{NRi} το διορθωμένο για μη ανταπόκριση βάρος έχουμε:

$$w_{NRi} = w_{Bi} RR_i^{-1}$$

4.1.3 Εκ των υστέρων στρωματοποίηση

Η τελευταία διόρθωση των βαρών γίνεται ώστε να συμφωνήσουν τα σύνολα στο δείγμα με γνωστές τιμές στον πληθυσμό. Η διαδικασία αυτή διορθώνει δύο σφάλματα. Εκείνο που οφείλεται στην μη-ανταπόκριση αλλά και εκείνο που οφείλεται στην μη κάλυψη του πληθυσμού. Αυτό έγινε με χρήση στοιχείων της Ελληνικής Στατιστικής Εταιρίας από την απογραφή του 2011 τα οποία αφορούν στην κατανομή του ελληνικού πληθυσμού ως προς ηλικία και φύλλο. Στα δεδομένα αυτά όπως δημοσιεύτηκαν οι ηλικιακές ομάδες είναι ομαδοποιημένες ανά δεκαετία (1-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65+). Επειδή η ΕΜΕΝΟ αφορά στον ενήλικο πληθυσμό η πρώτη κατηγορία δεν λήφθηκε υπόψη. Ανά ηλικία και φύλο, δημιουργήθηκαν 12 κατηγορίες. Έστω ότι ένα άτομο ανήκει στην $m_{j,k}$ κατηγορία με j,k το φύλλο και η ηλικιακή ομάδα αντίστοιχα. Τότε στο άτομο αυτό αντιστοιχεί ένας παράγοντας διόρθωσης w_{psa} ίσος με:

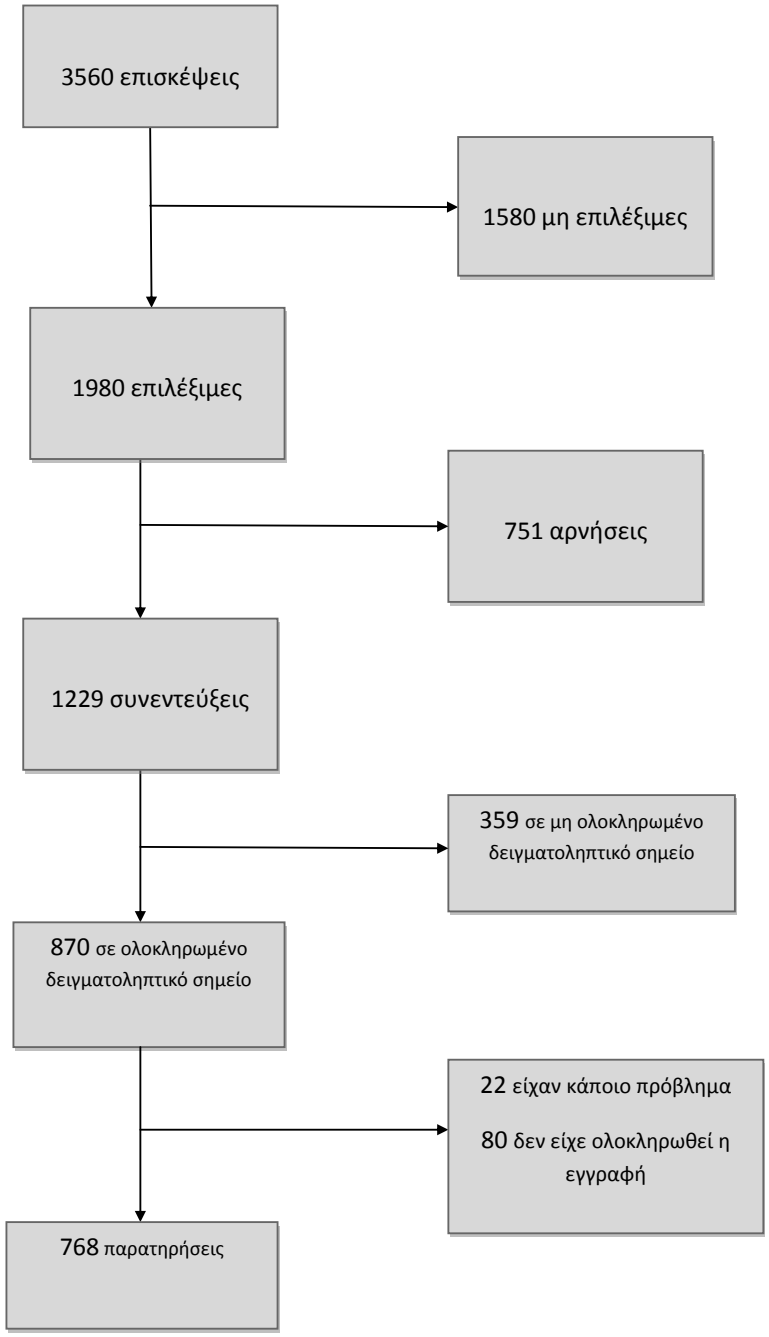
$$w_{psai} = \frac{N_m}{\sum_{i=1}^{n_m} w_{NRi}}$$

όπου w_{NRi} το διορθωμένο για μη ανταπόκριση βάρος, N_m το σύνολο του πληθυσμού στην $m_{j,k}$ κατηγορία και n_m το σύνολο των ανταποκριθέντων που ανήκουν στην $m_{j,k}$ κατηγορία.

4.2 Προετοιμασία των δεδομένων

4.2.1 Τα δεδομένα

Η ανάλυση των δεδομένων έγινε σε ένα υποσύνολο των δεδομένων του κυρίου μέρους της μελέτης ΕΜΕΝΟ. Λόγω πρακτικών περιορισμών, και ενώ η κύρια ανάλυση της μελέτης θα γίνει σε συνολικό δείγμα 6000 ατόμων, στην παρούσα φάση θα χρησιμοποιηθεί μέρος αυτών, εκείνα τα οποία είχαν καταγραφεί πλήρως μέχρι την 15/12/2014. Περιορισμός αποτελεί ο σχετικά μικρός αριθμός δεδομένων, διότι το δείγμα αυτό σε καμία περίπτωση δεν μπορεί να θεωρηθεί αντιπροσωπευτικό και τα αποτελέσματα δεν μπορούν να γενικευθούν. Με τα μέχρι στιγμής δεδομένα και εξαιρώντας από την ανάλυση μας παρατηρήσεις που δεν ανήκουν σε ολοκληρωμένο δειγματοληπτικό σημείο καταλήγουμε σε 768 άτομα. Ολοκληρωμένο θεωρείται ένα δειγματοληπτικό σημείο στο οποίο έχουν ολοκληρωθεί 12 συνεντεύξεις όταν πρόκειται για αστική η ημιαστική περιοχή, αντίστοιχα 8 συνεντεύξεις σε αγροτική περιοχή. Στο διάγραμμα ροής 4.1 φαίνεται πως καταλήγουμε στο δείγμα από το σύνολο κατοικιών που επισκέφθηκαν οι συνεργάτες της ΕΜΕΝΟ.Στον πίνακα 4.1 παρουσιάζονται τα αποτελέσματα του συνόλου των επισκέψεων που πραγματοποιήθηκαν. Λόγω του ότι το συνολικό ποσοστό ανταπόκρισης υπολογίζεται στο σύνολο των επιλέξιμων, τα αποτελέσματα των επισκέψεων μόνο για τις επιλέξιμες κατοικίες παρουσιάζονται στον πίνακα 4.2, από τον οποίο βλέπουμε ότι το συνολικό ποσοστό ανταπόκρισης ήταν 62.07%.



Σχήμα 4.1: Διάγραμμα ροής:Περιγραφή της διαδικασίας μέχρι το τελικό δείγμα. Αναλυτικά οι λόγοι άρνησεις και μη επιλεξιμότητας παρουσιάζονται στον πίνακα 4.1

Τελικό Αποτέλεσμα επίσκεψης	N (%)
Δευτερεύουσα κατοικία	47(1.3)
Ακατοίκητο	330(9.27)
Επαγγελματική χρήση	210(5.9)
Απουσιάζουν για περισσότερο από 2 εβδομάδες	299(8.4)
Απουσίαζαν και στις 3 επισκέψεις	673(18.9)
Απροσδιόριστο	21(0.59)
Πρόβλημα γλώσσας	41(1.15)
Αρνήθηκε: Δεν αιτιολόγησε	397(11.15)
Αρνήθηκε: Λόγω έλλειψης χρόνου	72(2.02)
Αρνήθηκε: Θέμα προσωπικής αρχής (Συνειδητά δε συμμετέχει σε έρευνες)	24(0.67)
Αρνήθηκε: Λόγω ασθενείας	28(0.79)
Αρνήθηκε: Αισθάνεται υγιής (έτσι δε νομίζει ότι έχει λόγο να συμμετάσχει)	26(0.73)
Αρνήθηκε: Το θέμα της έρευνας (δεν ενδιαφέρεται για θέματα υγείας ή τα θεωρεί πολύ προσωπικά)	35(0.98)
Αρνήθηκε για άλλο λόγο	128(3.6)
Συνέντευξη	1229(34.5)
Σύνολο	3560

Πίνακας 4.1: Αποτελέσματα τελευταίας επίσκεψης κατοικίας

Τελικό Αποτέλεσμα επίσκεψης	N (%)
Πρόβλημα γλώσσας	41(2.07)
Αρνήθηκε: Δεν αιτιολόγησε	397(20.05)
Αρνήθηκε: Λόγω έλλειψης χρόνου	72(3.64)
Αρνήθηκε: Θέμα προσωπικής αρχής (Συνειδητά δε συμμετέχει σε έρευνες)	24(1.21)
Αρνήθηκε: Λόγω ασθενείας	28(1.41)
Αρνήθηκε: Αισθάνεται υγιής (έτσι δε νομίζει ότι έχει λόγο να συμμετάσχει)	26(1.31)
Αρνήθηκε: Το θέμα της έρευνας (δεν ενδιαφέρεται για θέματα υγείας ή τα θεωρεί πολύ προσωπικά)	35(1.77)
Αρνήθηκε για άλλο λόγο	128(6.46)
Συνέντευξη	1229(62.07)
Σύνολο	1980

Πίνακας 4.2: Αποτελέσματα τελευταίας επίσκεψης κατοικίας εξαιρώντας τις μη επιλέξιμες.

4.3 Σκοπός και σχέδιο ανάλυσης

Ο σκοπός της ανάλυσης δεν είναι η εξαγωγή συμπερασμάτων για τον πληθυσμό. Είναι η εφαρμογή των μεθόδων και η σύγκριση των αποτελεσμάτων των τεσσάρων παρακάτω αναλύσεων:

1. Η πρώτη ανάλυση θα είναι η αδρή ανάλυση (naive analysis). Στην αδρή ανάλυση δεν λαμβάνεται υπόψιν ο σχεδιασμός της μελέτης, το δείγμα αντιμετωπίζεται ως απλό τυχαίο δείγμα, και τα βάρη αγνοούνται. Περιμένουμε αυτή η ανάλυση να δώσει μεροληπτικές εκτιμήσεις όσον αφορά τις σημειακές εκτιμήσεις, υποεκτιμημένα τυπικά σφάλματα και κατά συνέπεια, στενότερα διαστήματα εμπιστοσύνης.

2. Η δεύτερη ανάλυση θα λαμβάνει υπόψιν τον σχεδιασμό της μελέτης (design based analysis), ωστόσο θα αγνοεί την παρουσία μη ανταπόκρισης. Όλες οι τεχνικές εκτίμησης με βάση τον σχεδιασμό της μελέτης υποθέτουν πλήρη ανταπόκριση. Επομένως αναμένουμε οι εκτιμήσεις αυτές να περιέχουν μεροληψία.

3. Η τρίτη ανάλυση θα λαμβάνει υπόψιν τόσο τον σχεδιασμό της μελέτης όσο και την παρουσία μη ανταπόκρισης. Φυσικά θα ήταν υπέρ αισιόδοξο να πούμε ότι εξαλείφουμε το σφάλμα που οφείλεται στη μη ανταπόκριση. Αναμένουμε όμως σημαντική διόρθωση των εκτιμήσεων, ιδίως σε μεταβλητές οι οποίες σχετίζονται τόσο με την πιθανότητα ανταπόκρισης όσο και με τις μεταβλητές με βάση τις οποίες έγινε η διόρθωση. Στην περίπτωση της EMENO η διόρθωση έγινε ανά περιοχή, επομένως περιμένουμε να δούμε μεγαλύτερη διαφορά στις μεταβλητές εκείνες που σχετίζονται με την περιοχή.

4. Η τέταρτη ανάλυση θα έχει όλα τα χαρακτηριστικά της τρίτης ανάλυσης, επιπλέον όμως, θα περιέχει διόρθωση για την κατανομή του φύλου και της ηλικίας (εκ των υστέρων στρωματοποίηση). Αναμένουμε αυτή η ανάλυση να δίνει τις πιο αμερόληπτες εκτιμήσεις.

Κεφάλαιο 5

Αποτελέσματα

5.1 Περιγραφή του δείγματος

Αρχικά μας ενδιαφέρει να εξετάσουμε κάποια βασικά χαρακτηριστικά του δείγματος καθώς και το πως αλλάζουν οι εκτιμήσεις μας ανάλογα με την μέθοδο ανάλυσης. Το δείγμα μας αποτελείται από 768 άτομα με μέση ηλικία 56.3 (12.27) έτη και εύρος (18-101), ενώ οι 352 είναι άνδρες (45.83%) και οι 416 γυναίκες (54.17%). Οι περιοχές από τις οποίες προέρχονται τα άτομα καθώς και η μέση ηλικία ανά περιοχή φαίνονται στον πίνακα 5.1 ενώ στον πίνακα 5.2 η κατανομή του φύλου ανά περιοχή. Σύμφωνα με τους πίνακες 5.1 5.2 το δείγμα αποτελείται σε μεγαλύτερο ποσοστό από γυναίκες, και η μέση ηλικία είναι κάπως αυξημένη σε σχέση με εκείνη του πληθυσμού, η οποία είναι τα 41.9 έτη σύμφωνα με την απογραφή του 2011 (*Announcement of the demographic and social characteristics of the Resident Population of Greece according to the 2011 Population - Housing Census* 2013).

Στον πίνακα 5.4 φαίνονται βασικά χαρακτηριστικά. Τα στοιχεία αυτά αποτελούν περιγραφή του δείγματος. Παρακάτω θα εκτιμήσουμε τα ποσοστά αυτά με μεθόδους που λαμβάνουν υπόψιν τον σχεδιασμό της μελέτης, την μη ανταπόκριση αλλά και την σύνθεση του πληθυσμού.

Περιοχή	N (%)	Μέση ηλικία (SD)
Αθήνα	98(12.76)	54.177(17.81)
Θεσσαλονίκη	104(13.54)	54.82 (17.74)
Πελοπόννησος	316(41.15)	57.06(17.25)
Ήπειρος	131(17.06)	58.17(16.14)
Βοιωτία	119(15.49)	55.3(17.59)
Σύνολο	768	56.3(17.27)

Πίνακας 5.1: Μέση ηλικία (Σταθερή απόκλιση-SD) του δείγματος ανά περιοχή

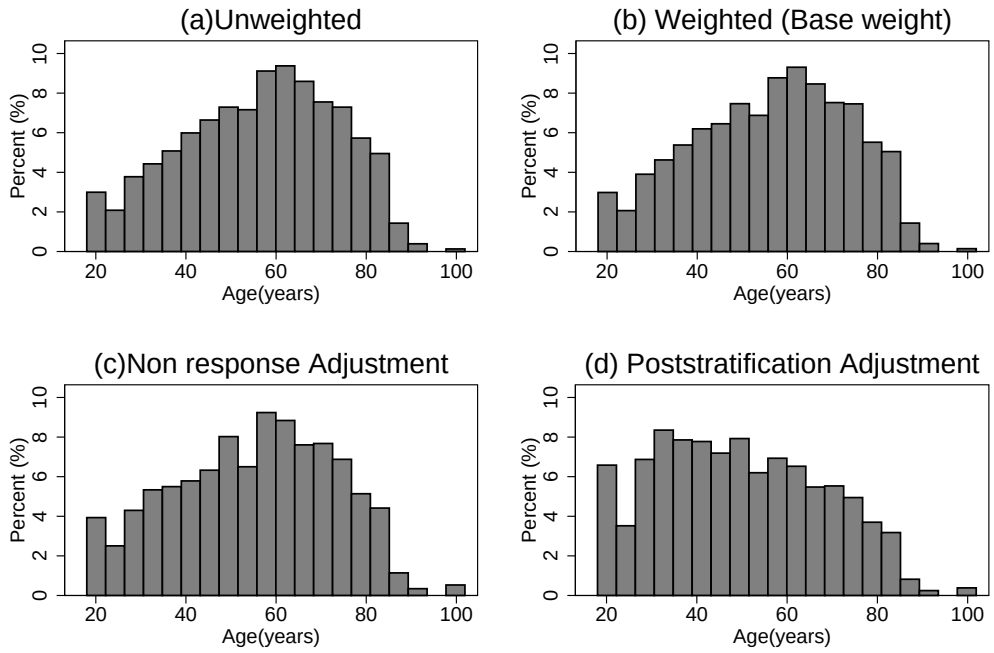
Περιοχή	Φύλο	N(%)	Σύνολο
Αθήνα	Άνδρας	41(41.84)	
	Γυναίκα	57(58.16)	98
Θεσσαλονίκη	Άνδρας	41(39.42)	
	Γυναίκα	63(60.58)	104
Πελοπόννησος	Άνδρας	163(51.58)	
	Γυναίκα	153(48.42)	316
Ήπειρος	Άνδρας	55(41.98)	
	Γυναίκα	76(58.02)	131
Βοιωτία	Άνδρας	52(43.70)	
	Γυναίκα	67(56.30)	119
	Σύνολο	768	

Πίνακας 5.2: Κατανομή φύλου του δείγματος ανά περιοχή

Στον πίνακα 5.3 βλέπουμε την κατανομή ηλικίας και φύλου του Ελληνικού πληθυσμού σύμφωνα με την απογραφή του 2011 σε σύγκριση με τις αντίστοιχες κατανομές του δείγματος. Παρατηρούμε ότι υπάρχει υποεκπροσώπηση των νέων κάτω των 40 και υπερεκπροσώπηση των άνω των 65. Για παράδειγμα ο πληθυσμός αποτελείται από 17.5% από άτομα ηλικίας 25-34, όμως το αντίστοιχο ποσοστό στο δείγμα είναι μόλις 9.51%. Αντίθετα η μεγαλύτερη ηλικιακή ομάδα 65+, η οποία αποτελεί το 24.12% του πληθυσμού, στο δείγμα βρίσκεται σε ποσοστό 35.81%. Επίσης υπάρχει υπερεκπροσώπηση των γυναικών, καθώς ο πληθυσμός αποτελείται κατά 51.38% από γυναίκες, όμως το αντίστοιχο ποσοστό του δείγματος είναι 54.17%. Λαμβάνοντας υπόψιν τα παραπάνω, κρίνεται απαραίτητη η διόρθωση ως προς ηλικία και φύλο ώστε η κατανομή του δείγματος να συμφωνεί με εκείνη του πληθυσμού. Στο σχήμα 5.1 φαίνεται η αλλαγή της κατανομής της ηλικίας σε κάθε στάδιο διόρθωσης των βαρών.

Ηλικιακή ομάδα	Ποσοστό(%) Πληθυσμός	Ποσοστό(%) Δείγμα
18-24	8.01	3.91
25-34	17.52	9.51
35-44	18.77	13.15
45-54	16.91	17.58
55-64	14.68	20.05
65+	24.12	35.81
Φύλο		
Άνδρας	48.62	45.83
Γυναίκα	51.38	54.17

Πίνακας 5.3: Σύγκριση κατανομής ηλικίας και φύλου του δείγματος με την αντίστοιχη του πληθυσμού όπως αυτή προκύπτει από την απογραφή του 2011



Σχήμα 5.1: Η κατανομή της ηλικίας του δείγματος υπολογισμένη:
(a):Χωρίς χρήση βαρών-Αδρή ανάλυση
(b):Με χρήση των αρχικών δειγματοληπτικών βαρών
(c):Με χρήση των διορθωμένων για μη ανταπόκριση βαρών
(d):Με χρήση των διορθωμένων βαρών για μη ανταπόκριση και για την από κοινού κατανομή ηλικίας-φύλου

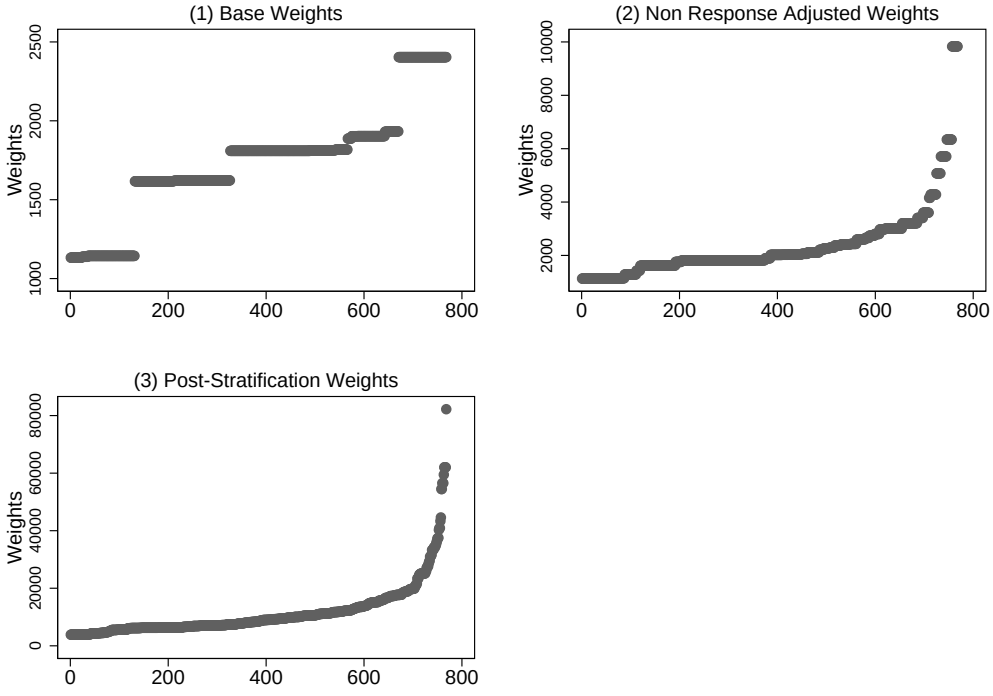
Ένα ακόμα στάδιο διόρθωσης που μπορεί να είναι απαραίτητο είναι η περικοπή των βαρών σε περίπτωση ύπαρξης ακραία μεγάλων τιμών. Γι αυτό το σκοπό, είναι απαραίτητη η γραφική εποπτεία των βαρών.

Στο σχήμα 5.2 παρουσιάζονται τα βάρη κάθε σταδίου, σε αύξουσα σειρά. Τα αρχικά-δειγματοληπτικά βάρη παίρνουν διακριτές τιμές οι οποίες είναι ορισμένες κατά τον σχεδιασμό. Τα διορθωμένα για μη ανταπόκριση βάρη δεν παίρνουν διακριτές τιμές όμως φαίνεται να δίνεται μεγάλο βάρος σε κάποιες παρατηρήσεις συγκριτικά με τις υπόλοιπες. Τέλος στα τελικά βάρη μόνο μια παρατήρηση παίρνει μεγάλο βάρος. Αυτό δεν θεωρείται μεγάλο πρόβλημα επομένως δεν θα προχωρήσουμε σε περικοπή των βαρών. Τέλος, παρατηρούμε ότι η κλίμακα των βαρών αυξάνεται σε κάθε στάδιο διόρθωσης. Αυτό συμβαίνει διότι σε κάθε στάδιο γίνεται πολλαπλασιασμός των βαρών με κάποιον παράγοντα μεγαλύτερο της μονάδας. Το άθροισμα των τελικών βαρών πρέπει να είναι ίσο με το σύνολο των ατόμων του πληθυσμού αναφοράς. Πολλές φορές γίνεται αλλαγή της κλίμακας των βαρών, ώστε αυτά να αθροίζουν σε κάποιο άλλο σύνολο. Συνήθως το σύνολο αυτό επιλέγεται να είναι το μέγεθος του δείγματος. Στην πραγματικότητα αυτό γίνεται για ευκολία και δεν έχει καμία επίδραση στις εκτιμήσεις. Στην παρούσα ανάλυση δεν έγινε αλλαγή της κλίμακας των βαρών και γι αυτό το λόγο κατά την εποπτεία των βαρών του σχήματος 5.2, σημασία έχει μόνο η μορφή των γραφημάτων, και όχι η αριθμητική τιμή των βαρών.

Δημογραφικά χαρακτηριστικά

Μεταβλητή	N(%)
Φύλο	
Άνδρας	352(45.83)
Γυναίκα	416(54.17)
Οικογενειακή Κατάσταση	
Άγαμος/η	140(18.23)
Έγγαμος/η	491(63.93)
Χήρος/α	90(11.72)
Διαζευγμένος/η	40(5.21)
Δεν Απαντώ	7(0.91)
Έγγαμος ή Συμβίωση	
Ναι	514(66.93)
Όχι	253(32.94)
Άγνωστο	1(0.13)
Παιδιά	
Όχι	163(21.22)
Ναι	591(76.95)
Άγνωστο	14(1.82)
Επίπεδο εκπαίδευσης	
Δεν έχω ολοκληρώσει το δημοτικό	91(11.85)
Δημοτικό	216 (28.13)
Γυμνάσιο και κατώτερες τεχνικές σχολές	93(12.11)
Λύκειο ή Τεχνικό Επαγγελματικό Εκπαιδευτήριο	192(25.00)
ΙΕΚ,ΚΕΚ, κολέγια	48(6.25)
ΤΕΙ	28(3.65)
ΑΕΙ	71(9.24)
Μεταπτυχιακό	10(1.30)
Διδακτορικό Δίπλωμα	5(0.65)
Άγνωστο	14(1.82)
Εργασιακή Κατάσταση	
Απασχολούμενος/η	274(35.68)
Άνεργος/η	92(11.98)
Μαθητής, σπουδαστής, φοιτητής	19(2.47)
Συνταξιούχος	286(37.24)
Οικιακά	73(9.51)
Μη εργαζόμενος (Αναπηρία)	9(1.17)
Στρ. Θητεία	1(0.13)
Άλλο	10(1.30)
Άγνωστο	4(0.52)
Εισόδημα	
<900 €	324(45.83)
900-1700€	202(28.57)
>1700 €	56(7.92)
Άγνωστο	125(17.68)

Πίνακας 5.4: Δημογραφικά χαρακτηριστικά δείγματος



Σχήμα 5.2: Τα βάρη της ανάλυσης, ταξινομημένα με αύξουσα σειρά. Στον οριζόντιο άξονα φαίνεται η κατάταξη των βαρών.
(1)Αρχικά
(2)Διορθωμένα για μη ανταπόκριση
(3)Διορθωμένα για μη ανταπόκριση και για την κατανομή ηλικίας-φύλου σύμφωνα με την αντίστοιχη του πληθυσμού

5.2 Εκτιμήσεις βασικών χαρακτηριστικών

Τα ποσοστά που παρουσιάστηκαν στον πίνακα 5.4 είναι απλά περιγραφικά του δείγματος και όχι εκτιμήσεις για τον πληθυσμό. Για να κάνουμε εκτιμήσεις χρειάζεται να λάβουμε υπόψιν τον σχεδιασμό της μελέτης. Στον πίνακα 5.5 επαναλαμβάνουμε τα αποτελέσματα του πίνακα 5.4, όπως προκύπτουν από ανάλυση με βάση τον σχεδιασμό με χρήση των διαφορετικών βαρών από το κάθε στάδιο διόρθωσης.

Ενδιαφέρον έχει το πως τροποποιείται η εκτίμηση του ποσοστού για τους εργαζόμενους και τους συνταξιούχους καθώς στην πρώτη και δεύτερη στήλη τα ποσοστά είναι σχεδόν ίδια (36.1% και 37.40% αντίστοιχα) ενώ στην τρίτη αλλάζουν (43.70% και 26.40% αντίστοιχα). Στο δείγμα έχουμε καθαρή υπερεκπροσώπηση ατόμων μεγαλύτερων ηλικιακών ομάδων που πιο πιθανό είναι να είναι συνταξιούχοι υπερεκτιμώντας έτσι το ποσοστό. Διορθώνοντας για μη ανταπόκριση αυτό δεν διορθώνεται καθώς η διόρθωση αυτή δεν σχετίζεται με την ηλικία. Διορθώνοντας όμως και για την κατανομή της ηλικίας, η εκτίμηση αλλάζει καθώς στους νεότερους ανατίθενται τελικά μεγαλύτερα βάρη. Το ίδιο ισχύει και για την εκτίμηση της μέσης ηλικίας. Κάνοντας χρήση των δύο πρώτων σετ βαρών, η εκτίμηση της μέσης ηλικίας είναι μεροληπτική. Αντίθετα η τελευταία εκτίμηση είναι πολύ κοντά στην πραγματική μέση τιμή.

Μεταβλητή	Base weights % (95% ΔΕ)	NR weights %(95% ΔΕ)	PS weights %(95% ΔΕ)
Φύλο			
Άντρας	45.8(40.8,51.00)	43.4(36.00,50.00)	48.5(41.50,55.50)
Γυναίκα	54.2(49.00,59.20)	56.6(50.00,63.10)	51.50(44.50,58.50)
Οικογενειακή Κατάσταση			
Άγαμος/η	18.40(14.30,23.20)	21.40(16.00,28.00)	31.00(23.50,39.60)
Έγγαμος/η	63(58.50,67.20)	59.60(53.40,65.60)	53.60(46.60,60.40)
Χήρος/α	11.90(8.92,15.70)	11.70(8.50,15.90)	8.66(6.08,12.20)
Διαζευγμένος	5.73(4.25,7.68)	6.44(5.02,8.23)	5.88(4.47,7.70)
Δεν απαντώ	1.01(0.39,2.6)	0.81(0.31,2.12)	0.88(0.32,0.43)
Έγγαμος ή συμβίωση			
Ναι	66.10(61.90,70.00)	64.40(59.00,69.40)	59.10(51.40,66.40)
Όχι	33.90(30.00,38.10)	35.60(30.60,41.00)	40.90(33.60,48.60)
Παιδιά			
Όχι	22.00(18.20,26.30)	25.00(20.30,30.30)	33.30(27.90,39.10)
Ναι	78.00(73.70,81.80)	75.00(69.70,79.70)	66.70(60.90,72.10)
Επίπεδο εκπαίδευσης			
Δεν έχω ολοκληρώσει το δημοτικό	12.00(9.17,15.50)	10.60(7.89,14.00)	8.19(6.02,11.00)
Δημοτικό	26.80(22.90,31.00)	23.80(20.30,27.60)	19.50(16.30,23.20)
Γυμνάσιο και κατώτερες τεχνικές σχολές	12.10(9.64,15.10)	10.50(8.09,13.50)	10.20(7.86,13.20)
Λύκειο ή ΤΕΕ	26.10(22.20,30.30)	27.70(22.30,33.80)	30.60(23.60,38.70)
ΙΕΚ,ΚΕΚ, κολέγια	6.63(4.67,9.33)	6.94(5.02,9.52)	8.31(5.88,11.60)
ΤΕΙ	4.07(2.54,6.46)	4.29(2.64,6.91)	5.60(3.19,9.65)
ΑΕΙ	10.20(7.58,13.50)	12.50(9.19,16.90)	13.50(9.15,19.50
Μεταπτυχιακό	1.58(0.71,3.48)	2.02(0.96,4.19)	2.52(1.12,5.60)
Διδακτορικό	0.67(0.25,1.73)	1.66(0.67,4.01)	1.51(0.56,3.95)
Εργασιακή κατάσταση			
Απασχολούμενος/η	36.1(30.90,41.60)	36.40(31.20,42.00)	43.70(37.40,50.10)
Άνεργος/η	12.00(9.74,14.80)	12.80(10.40,15.70)	14.60(11.50,18.30)
Μαθητής, σπουδαστής, φοιτητής	2.62(1.11,6.05)	3.86(1.06,13.10)	6.22(1.52,22.20)
Συνταξιούχος	37.40(32.10,43.10)	36.30(30.20,42.90)	26.40(20.90,32.80)
Οικιακά	9.18(6.96,12.00)	8.08(5.95,10.90)	6.79(4.85,9.40)
Μη εργαζόμενος (Αναπηρία)	1.19(0.55,2.55)	1.02(0.46,2.23)	0.76(0.35,1.65)
Στρ. Θητεία	0.136(0.01,1.39)	0.11(0.01,1.16)	0.24(0.02,2.46)
Άλλο	1.38(0.68,2.76)	1.39(0.68,2.79)	1.34(0.64,2.75)
Εισόδημα			
< 900 €	54.36(47.81,60.76)	51.73(43.86,59.52)	51.4(41.04,61.64)
900-1700€	35.31(30.15,40.83)	35.8(30.67,41.27)	36.22(29.56,43.46)
> 1700 €	10.33(7.11,14.77)	12.47(8.05,18.82)	12.38(7.71,19.28)
Μέση ηλικία	56.14(53.87,58.41)	55.02(52.13,57.91)	49.57(46.01,53.13)

Πίνακας 5.5: Ανάλυση με βάση τον σχεδιασμό: Εκτιμήσεις και σύγκριση των βασικών χαρακτηριστικών που προκύπτουν με χρήση διαφορετικών βαρών:
Base weights: Ανάλυση με τα δειγματοληπτικά βάρη
NR weights: Ανάλυση με τα διορθωμένα για μη ανταπόκριση βάρη
PS weights : Ανάλυση με τα διορθωμένα βάρη για μη ανταπόκριση και για κατανομή ηλικίας-φύλου.

5.3 Γνώσεις και στάσεις σχετικά με τον HIV

Στο πλαίσιο της Hprolipsis το ερωτηματολόγιο περιείχε πρωτυποποιημένες ερωτήσεις σχετικά με τις γνώσεις και στάσεις αναφορικά με τα λοιμώδη νοσήματα ηπατίτιδα Β (HBV) ηπατίτιδα C (HCV)και AIDS (HIV) . Αυτές οι ερωτήσεις είναι σχετικές με τα χαρακτηριστικά του νοσήματος, την αντίστοιχη θεραπεία και τους πιθανούς τρόπους μετάδοσης. Οι πιθανές απαντήσεις είναι: Σωστό, Λάθος, Δεν ξέρω, Δεν απαντώ. Δημιουργήθηκαν τρία σκόρ ως εξής:

1. Σκορ γνώσεων: Συν ένας βαθμός για κάθε σωστή ερώτηση η οποία ορθώς απαντήθηκε ¨Σωστή¨. Έτσι υψηλό σκορ γνώσεων σημαίνει υψηλή κατανόηση. Κατώτερη δυνατή τιμή του σκορ είναι το μηδέν ενώ ανώτατη το οκτώ.

2. Σκορ παρανοήσεων: Συν ένας βαθμός για κάθε λανθασμένη ερώτηση που λανθασμένα δεν απαντήθηκε ¨Λάθος¨. Έτσι υψηλό σκορ παρανοήσεων σημαίνει υψηλή παρανόηση. Κατώτερη δυνατή τιμή του σκορ είναι το μηδέν ενώ ανώτατη το δέκα.

3. Συνολικό επίπεδο γνώσεων: Η διαφορά των δύο παραπάνω σκορ (Γνώσεις-Παρανοήσεις). Έτσι θετικό σκορ δηλώνει περισσότερες γνώσεις ενώ αρνητικό περισσότερες παρανοήσεις. Κατώτερη δυνατή τιμή του συνολικού σκορ είναι το μείον δέκα ενώ ανώτατη το οκτώ.

Στη συγκεκριμένη διπλωματική εργασία, θα ασχοληθούμε μόνο με τις ερωτήσεις σχετικά με το AIDS (HIV). Οι ερωτήσεις οι οποίες έγιναν συνοδευόμενες από την σωστή απάντηση παρουσιάζονται στον πίνακα 5.3 και οι απαντήσεις οι οποίες δόθηκαν, παρουσιάζονται στον πίνακα 5.7. Οι απαντήσεις Δεν ξέρω και Δεν απαντώ θεωρήθηκαν ίδιες εννοιολογικά επομένως παρουσιάζονται ως μια κατηγορία. Τα ποσοστά Δεν ξέρω-Δεν απαντώ είναι εντυπωσιακά υψηλά καθώς κυμαίνονται από 24.3%-62.4%, με το ποσοστό στις περισσότερες ερωτήσεις να είναι κοντά στο 50%.

Ερωτήσεις με το υψηλότερο ποσοστό σωστών απαντήσεων:

- Το 74.3% των ερωτηθέντων απάντησε ότι το HIV/AIDS μεταδίδεται μέσω μετάγγισης από μολυσμένο αίμα.
- Το 70% των ερωτηθέντων απάντησε ότι το HIV/AIDS μεταδίδεται μέσω σεξουαλικής επαφής χωρίς προφυλακτικό.
- Το 62.5% των ερωτηθέντων απάντησε ότι το HIV/AIDS μεταδίδεται με χρήση ενδοφλεβίων ναρκωτικών.

Ερωτήσεις με το υψηλότερο ποσοστό λανθασμένων απαντήσεων:

- Το 31.7% των ερωτηθέντων απάντησε ότι υπάρχει οριστική θεραπεία για το HIV/ AIDS.
- Το 36.9% των ερωτηθέντων απάντησε ότι το HIV και το AIDS είναι το ίδιο.
- Το 29.4% των ερωτηθέντων απάντησε ότι το HIV/ AIDS μεταδίδεται από τσίμπημα κουνουπιού.

Ερώτηση	Σωστή Απάντηση
Το HIV και το AIDS είναι το ίδιο	Λάθος
Όταν κάποιος έχει HIV έχει απαραίτητα συμπτώματα	Λάθος
Αν κάποιος έχει HIV/AIDS αλλά φαίνεται και νοιώθει υγιής, τότε δεν μπορεί να μεταδωσει τον ιό	Λάθος
Η HIV λοίμωξη μπορεί να προκαλέσει πτώση της άμυνας του οργανισμού	Σωστό
Υπάρχει εμβόλιο για το HIV/ AIDS	Λάθος
Ο μόνος τρόπος για να μάθει κάποιος/α εάν είναι θετικός/η είναι να κάνει το αντίστοιχο τεστ	Σωστό
Υπάρχει θεραπεία για το HIV/ AIDS	Σωστό
Υπάρχει οριστική θεραπεία για το HIV/ AIDS	Λάθος
Με ποιους από τους παρακάτω τρόπους μεταδίδεται το HIV/ AIDS ·	
Μετάγγιση από μολυσμένο αίμα	Σωστό
Σεξουαλική επαφή χωρίς προφυλακτικό	Σωστό
Από θετική μητέρα στο έμβρυο	Σωστό
Καθημερινή κοινωνική επαφή (χειραψία, συνομιλία)	Λάθος
Με το να πίνεις ή να τρως από τα ίδια σκεύη με κάποιον που είναι θετικός	Λάθος
Με το να χρησιμοποιείς την ίδια τουαλέτα, πισίνα, σάουνα με κάποιον που είναι θετικός	Λάθος
Φιλί	Λάθος
Τσίμπημα κουνουπιού	Λάθος
Τατουάζ ή Τρυπήματα του σώματος (Body piercing)	Σωστό
Χρήση ενδοφλεβίων ναρκωτικών	Σωστό

Πίνακας 5.6: Σωστές απαντήσεις στις ερωτήσεις σχετικά με τον HIV/ AIDS .

Ερώτηση	Απάντηση	(N%)
Το HIV και το AIDS είναι το ίδιο.		
	Σωστό	261 (36.9)
	Λάθος	57 (8.1)
	Δεν Απαντώ	388 (54.9)
	Άγνωστο	1 (0.1)
Όταν κάποιος έχει HIV έχει απαραίτητα συμπτώματα.		
	Σωστό	188 (26.6)
	Λάθος	142 (20.1)
	Δεν Απαντώ	369 (52.2)
	Άγνωστο	8 (1.1)
Αν κάποιος έχει HIV/AIDS αλλά φαίνεται και νοιώθει υγιής, τότε δεν μπορεί να μεταδωσει τον ιό.		
	Σωστό	113 (16.0)
	Λάθος	221 (31.3)
	Δεν Απαντώ	367 (51.9)
	Άγνωστο	6 (0.8)
Η HIV λοίμωξη μπορεί να προκαλέσει πτώση της άμυνας του οργανισμού.		
	Σωστό	332 (47.0)
	Λάθος	17 (2.4)
	ΔΞ-ΔΑ	351 (49.6)
	Άγνωστο	7 (1.0)
Υπάρχει εμβόλιο για το HIV/AIDS.		
	Σωστό	110 (15.6)
	Λάθος	142 (20.1)
	Δεν Απαντώ	441 (62.4)
	Άγνωστο	14 (2.0)
Ο μόνος τρόπος για να μάθει κάποιος/α εάν είναι θετικός/η είναι να κάνει το αντίστοιχο τεστ.		
	Σωστό	158 (22.3)
	Λάθος	117 (16.5)
	ΔΞ-ΔΑ	423 (59.8)
	Άγνωστο	9 (1.3)
Υπάρχει θεραπεία για το HIV/AIDS.		
	Σωστό	281 (39.7)
	Λάθος	41 (5.8)
	Δεν Απαντώ	360 (50.9)
	Άγνωστο	25 (3.5)
Υπάρχει οριστική θεραπεία για το HIV/AIDS.		
	Σωστό	224 (31.7)
	Λάθος	118 (16.7)
	Δεν Απαντώ	358 (50.6)
	Άγνωστο	7 (1.0)
Με ποιους από τους παρακάτω τρόπους μεταδίδεται το HIV/AIDS.		
Μετάγγιση από μολυσμένο αίμα.		
	Σωστό	525 (74.3)
	Λάθος	2 (0.3)
	Δεν Απαντώ	172 (24.3)
	Άγνωστο	8 (1.1)
Σεξουαλική επαφή χωρίς προφυλακτικό.		
	Σωστό	495 (70.0)

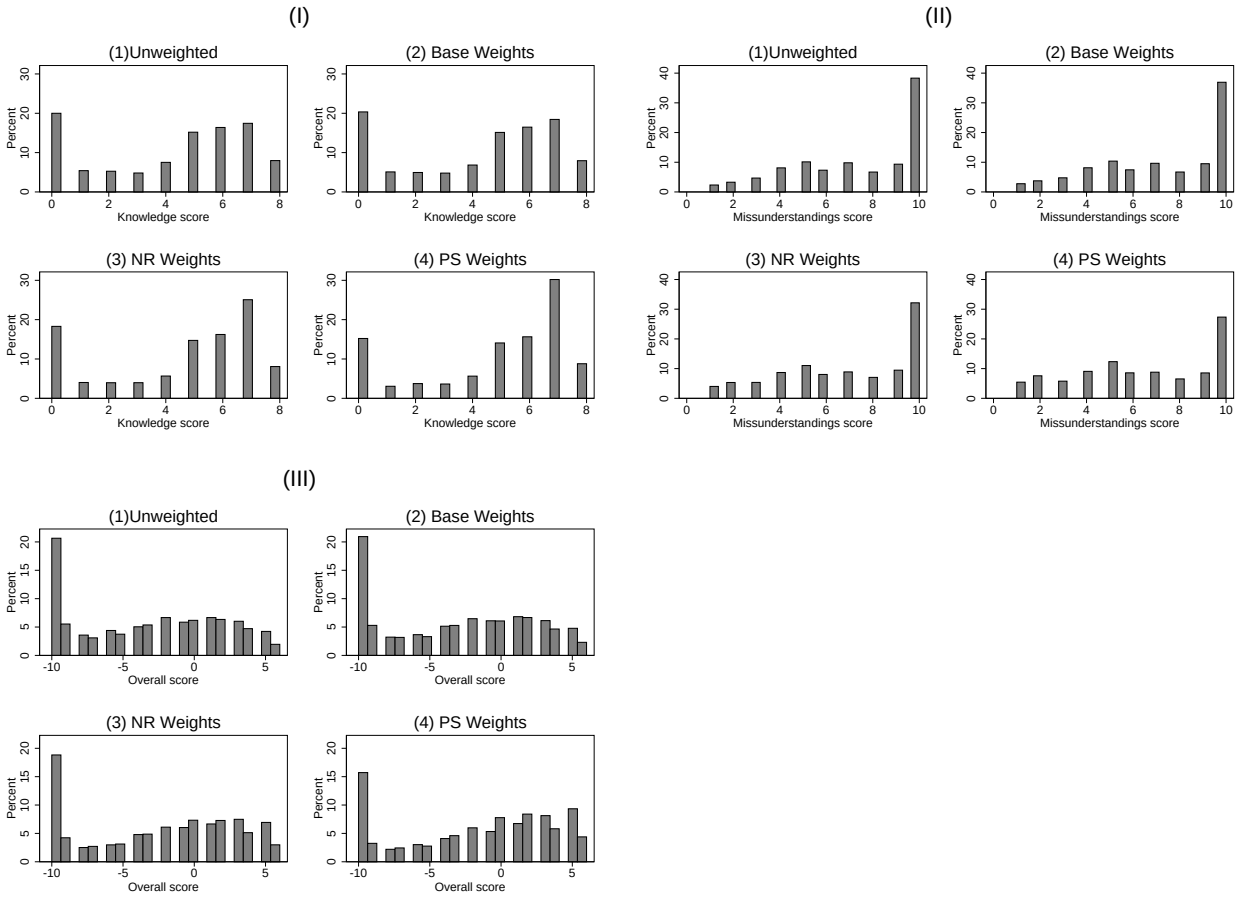
	Λάθος	3 (0.4)
	Δεν Απαντώ	200 (28.3)
	Άγνωστο	9 (1.3)
Από θετική μητέρα στο έμβρυο.		
	Σωστό	330 (46.7)
	Λάθος	90 (12.7)
	Δεν Απαντώ	278 (39.3)
	Άγνωστο	9 (1.3)
Καθημερινή κοινωνική επαφή (χειραψία, συνομιλία.)		
	Σωστό	64 (9.1)
	Λάθος	353 (49.9)
	Δεν Απαντώ	274 (38.8)
	Άγνωστο	16 (2.3)
Με το να πίνεις ή να τρως από τα ίδια σκεύη με κάποιον που είναι θετικός.		
	Σωστό	104 (14.7)
	Λάθος	257 (36.4)
	Δεν Απαντώ	328 (46.4)
.	Άγνωστο	18 (2.5)
Με το να χρησιμοποιείς την ίδια τουαλέτα, πισίνα, σάουνα με κάποιον που είναι θετικός.		
	Σωστό	134 (19.0)
	Λάθος	216 (30.6)
	Δεν Απαντώ	345 (48.8)
	Άγνωστο	12 (1.7)
Φιλί		
	Σωστό	175 (24.8)
	Λάθος	189 (26.7)
	Δεν Απαντώ	326 (46.1)
	Άγνωστο	17 (2.4)
Τσίμπημα κουνουπιού.		
	Σωστό	208 (29.4)
	Λάθος	135 (19.1)
	Δεν Απαντώ	347 (49.1)
	Άγνωστο	17 (2.4)
Τατουάζ ή Τρυπήματα του σώματος (Body piercing).		
	Σωστό	368 (52.1)
	Λάθος	63 (8.9)
	Δεν Απαντώ	263 (37.2)
	Άγνωστο	9 (1.3)
Ξρήση ενδοφλεβίων ναρκωτικών.		
	Σωστό	442 (62.5)
	Λάθος	29 (4.1)
	Δεν Απαντώ	229 (32.4)
	Άγνωστο	7 (1.0)

Πίνακας 5.7: Απαντήσεις που δόθηκαν στις ερωτήσεις σχετικά με τις γνώσεις για τον HIV/AIDS. Αδρή ανάλυση.

Στον πίνακα 5.8 παρουσιάζονται οι μέσες τιμές των τριών σκορ όπως προέκυψαν από τις 4 αναλύσεις. Για το σκορ γνώσεων η μέση τιμή είναι 4.2 στην πρώτη στήλη ενώ αλλάζει σημαντικά σε 4.89 κατά την τέταρτη. Όπως φαίνεται και από τα ιστογράμματα του σκορ γνώσεων 5.3 σύμφωνα με τις δύο πρώτες αναλύσεις η πλειοψηφία έχει σκορ γνώσεων ίσο με μηδέν, όμως σύμφωνα με την τρίτη και τέταρτη ανάλυση η πλειοψηφία έχει σκορ ίσο με επτά, με αποτέλεσμα να αυξάνεται η μέση τιμή. Παρομοίως η μέση τιμή για τα άλλα δύο σκορ αλλάζει στις δύο τελευταίες απαντήσεις. Η μέση τιμή του σκορ παρανοήσεων μειώνεται από 7.3 σε 6.5 με την πλειοψηφία να έχει σκορ ίσο με δέκα σε κάθε μια από τις αναλύσεις. Τέλος το μέσο συνολικό επίπεδο γνώσεων αυξάνεται από -3.7 σε -1.5, η πλειοψηφία έχει επίπεδο γνώσεων ίσο με -10 σε κάθε ανάλυση.

	Naive analysis Μέση τιμή (ΤΑ)	Base weights Μέση τιμή (95%ΔΕ)	NR weights Μέση τιμή (95%ΔΕ)	PS weights Μέση τιμή (95%ΔΕ)
Σκορ Γνώσεων	4.2 (2.75)	4.23 (3.87,4.59)	4.57 (4.24,4.893)	4.89 (4.56,5.22)
Σκορ παρανοήσεων	7.39 (2.712)	7.295 (6.95,7.63)	6.945 (6.56,7.32)	6.517 (6.03,7.003)
Συνολικό Σκορ	-3.178 (5.13)	-3.040 (-3.73,-2.35)	-2.349 (-3.05,-1.644)	-1.568 (-2.40,-0.73)

Πίνακας 5.8: Μέσο Σκορ γνώσεων, παρανοήσεων και συνολικό επίπεδο γνώσεων, σχετικά με τις γνώσεις για τον HIV .
Naive analysis: Ανάλυση η οποία αγνοεί τον σχεδιασμό
Base weights: Ανάλυση με τα δειγματοληπτικά βάρη
NR weights: Ανάλυση με τα διορθωμένα για μη ανταπόκριση βάρη
PS weights : Ανάλυση με τα διορθωμένα βάρη για μη ανταπόκριση και για κατανομή ηλικίας-φύλου.
ΤΑ: Τυπική Απόκλιση
ΔΑ: Διάστημα εμπιστοσύνης



Σχήμα 5.3: Ιστογράμματα του επιπέδου γνώσεων σχετικά με τον HIV
(I):Σκορ γνώσεων σχετικά με τον HIV
(II):Σκορ παρανοήσεων σχετικά με τον HIV
(III):Συνολικό επίπεδο γνώσεων σχετικά με τον HIV
(1)Unweighted: Ανάλυση η οποία αγνοεί τον σχεδιασμό
(2)Base weights: Ανάλυση με τα δειγματοληπτικά βάρη
(3)NR weights: Ανάλυση με τα διορθωμένα για μη ανταπόκριση βάρη
(4)PS weights : Ανάλυση με τα διορθωμένα βάρη για μη ανταπόκριση και για κατανομή ηλικίας-φύλου.

5.3.1 Πολυπαραγοντικό μοντέλο για το συνολικό επίπεδο γνώσεων
Ανάλυση Πλήρων Παρατηρήσεων

Στην συνέχεια θα ασχοληθούμε μόνο με το συνολικό επίπεδο γνώσεων. Μας ενδιαφέρει η συσχέτιση του επιπέδου γνώσεων με την ηλικία, το φύλο, το επίπεδο εκπαίδευσης και το εισόδημα. Οι παρατηρήσεις με τυχόν ελλείπουσες τιμές σε κάποια από αυτές τις μεταβλητές ή στο συνολικό επίπεδο γνώσεων εξαιρούνται. Αυτή η ανάλυση ονομάζεται Ανάλυση Πλήρων Παρατηρήσεων (ΑΠΠ). Τα αποτελέσματα από την γραμμική παλινδρόμηση, οι συντελεστές και αντίστοιχα 95% Διαστήματα Εμπιστοσύνης, φαίνονται στον πίνακα 5.9, και τα σχετικά διαγνωστικά γραφήματα παρουσιάζονται στο παράρτημα. Τα συμπεράσματα της αδρής ανάλυσης, η οποία αγνοεί τον σχεδιασμό, είναι τα εξής:

- Το φύλο δεν συσχετίζεται με το συνολικό επίπεδο γνώσεων.
- Υπάρχει ισχυρά στατιστικά σημαντική συσχέτιση μεταξύ ηλικίας και επιπέδου γνώσεων. Οι νεότεροι είναι περισσότερο ενήμεροι σχετικά με το HIV/AIDS . Συγκρίνοντας δύο άτομα με 5 χρόνια διαφορά ο νεότερος έχει περίπου μισή μονάδα υψηλότερο σκορ, διατηρώντας σταθερό το εκπαιδευτικό επίπεδο, το εισόδημα και το φύλο.
- Το επίπεδο εκπαίδευσης συσχετίζεται με το συνολικό επίπεδο γνώσεων επίσης ισχυρά στατιστικά σηματικά. Σε σύγκριση με άτομα των οποίων το επίπεδο εκπαίδευσης είναι μέχρι το δημοτικό, τα άτομα των οποίων η εκπαίδευση είναι μέχρι το Λύκειο αναμένουμε να έχουν περίπου 2.5 μονάδες υψηλότερο σκορ, ενω αναμένουμε 4.7 μονάδες υψηλότερο σκορ για τα άτομα των οποίων η εκπαίδευση είναι πάνω από Λύκειο.
- Το εισόδημα επίσης σχετίζεται με το σκορ. Τα άτομα των οποίων το εισόδημα ξεπερνάει τα 900 ευρώ αναμένουμε να έχουν 1.6 μονάδες υψηλότερο σκορ, σε σύγκριση με τα άτομα των οποίων το εισόδημα υπερβαίνει τα 900 ευρώ.

Αντίστοιχα είναι τα συμπεράσματα της ανάλυσης με βάση τον σχεδιασμό. Ανάλογα με τη μέθοδο στάθμισης υπάρχουν κάποιες διαφοροποιήσεις. Τα συμπεράσματα σχετικά με την συσχέτιση του φύλου και της ηλικίας με το επίπεδο γνώσεων, δεν αλλάζουν ανάλογα με την μέθοδο. Αυτό δεν συμβαίνει με το επίπεδο εκπαίδευσης και το εισόδημα. Οι συντελεστές του εκπαιδευτικού επιπέδου αυξάνουν, ιδιαίτερα όταν γίνεται χρήση των τελικών βαρών, και παραμένουν ισχυρά στατιστικά σημαντικοί (p-value < 0.001) ανεξάρτητα από τη μέθοδο στάθμισης. Αντίθετα η συσχέτιση του εισοδήματος με το επίπεδο γνώσεων παύει να είναι στατιστικά σημαντική όταν γίνει χρήση των τελικών βαρών.

Τα αποτελέσματα της τέταρτης στήλης θεωρούνται τα πιο αμερόληπτα καθώς η ανάλυση έχει γίνει με βάση τον σχεδιασμό και τα βάρη είναι πλήρως διορθωμένα για μη ανταπόκριση ατόμου. Γι αυτό το λόγο όλες οι επόμενες αναλύσεις θα γίνουν με χρήση αυτών των βαρών.

5.3.2 Ανάλυση με Πολλαπλές αντικαταστάσεις

Παρόλο που τα αποτελέσματα της τέταρτης στήλης του πίνακα 5.9 θεωρήθηκαν τα πιο αμερόληπτα, η παραπάνω ανάλυση δεν λάμβανε υπόψιν την μη ανταπόκριση ερώτησης (item nonresponse). Μη ανταπόκριση ερώτησης μπορεί να προκύψει τόσο στην μεταβλητή απόκρισης (συνολικό επίπεδο γνώσεων) όσο και σε κάποιες από τις επεξηγηματικές μεταβλητές. Το συνολικό επίπεδο γνώσεων προκύπτει ως άθροισμα από επιμέρους μεταβλητές. Έτσι, εάν σε κάποια από τις μεταβλητές αυτές δεν υπάρχει παρατηρηθείσα απάντηση, δεν είναι δυνατό να υπολογιστεί το άθροισμα. Συγκεκριμένα στο συνολικό επίπεδο γνώσεων υπάρχουν 88 (12.87%) ελλείπουσες τιμές, στο εισόδημα υπάρχουν 116 (16.96%) ελλείπουσες τιμές και στο επίπεδο εκπαίδευσης 14 (2.05%) ελλείπουσες τιμές.

Μεταβλητή	Naive analysis Συντελεστής 95% ΔΕ	Base weights Συντελεστής 95% ΔΕ	NR weights Συντελεστής 95% ΔΕ	PS weights Συντελεστής 95% ΔΕ
Ηλικία				
Ανά έτος:	-0.088**	-0.091**	-0.087**	-0.087**
	(-0.112,-0.063)	(-0.126,-0.055)	(-0.131,-0.042)	(-0.136,-0.038)
Φύλο				
Άνδρας	0			
Γυναίκα	0.023	0.030	-0.077	-0.302
	(-0.704,0.750)	(-0.891,0.950)	(-0.982,0.828)	(-1.260,0.656)
Επίπεδο εκπαίδευσης				
Μέχρι Δημοτικό	0			
Μέχρι Λύκειο	2.662 **	2.589 **	3.061**	3.274**
	(1.725,3.599)	(1.477,3.701)	(1.821,4.302)	(1.954,4.594)
Πάνω από Λύκειο	4.695**	4.463 **	4.889 **	4.956**
	(3.582,5.808)	(2.920,6.006)	(3.429,6.350)	(3.387,6.526)
Εισόδημα				
<900 €	0			
900-1700€	1.661 **	1.626*	1.353 *	1.214
	(0.852,2.471)	(0.538,2.714)	(0.237,2.470)	(-0.104,2.531)
>1700 €	1.548 *	1.204	1.551	1.685
	(0.222,2.875)	(-0.990,3.399)	(-0.595,3.696)	(-0.510,3.880)
Σταθερά	-0.860	-0.540	-0.601	-0.544
	(-2.704,0.983)	(-3.295,2.214)	(-4.146,2.943)	(-4.342,3.255)

Πίνακας 5.9: Πολλαπλή γραμμική παλινδρόμηση με εξαρτημένη μεταβλητή το συνολικό επίπεδο γνώσεων. Σύγκριση των σημειακών εκτιμήσεων και των διαστημάτων εμπιστοσύνης ανάλογα με τη μέθοδο στάθμισης:
(1)Naive analysis: Ανάλυση η οποία αγνοεί τον σχεδιασμό
(2)Base weights: Ανάλυση με τα δειγματοληπτικά βάρη
(3)NR weights: Ανάλυση με τα διορθωμένα για μη ανταπόκριση βάρη
(4)PS weights : Ανάλυση με τα διορθωμένα βάρη για μη ανταπόκριση και για κατανομή ηλικίας-φύλου.
* p-value < 0.05
** p-value < 0.001

Για την αντιμετώπιση του προβλήματος αυτού εφαρμόστηκαν μέθοδοι πολλαπλών αντικαταστάσεων (ΠΑ) με χρήση αλυσιδωτών εξισώσεων. Οι μεταβλητές εισόδημα και επίπεδο εκπαίδευσης είναι κατηγορικές με τρεις κατηγορίες και, ξεκάθαρα, υπάρχει διάταξη. Για τον λόγο αυτό οι μεταβλητές αυτές μοντελοποιήθηκαν με χρήση διαβαθμισμένης λογιστικής παλινδρόμησης (ordinal logistic). Η μεταβλητή του συνολικού επιπέδου γνώσεων αντιμετωπίζεται ως συνεχής παίρνει όμως διακριτές τιμές και δεν ακολουθεί κανονική κατανομή. Για αυτούς τους λόγους δεν έγινε χρήση γραμμικής παλινδρόμησης στο μοντέλο αντικατάστασης. Προτιμότερο θεωρήθηκε να γίνει χρήση της μεθόδου predictive mean matching (pmm).Όπως αναφέρθηκε στην ενότητα 3.1.1, η μέθοδος αυτή είναι μια hot deck μέθοδος αντικατάστασης. Γίνεται λοιπόν αντικατάσταση των ελλειπουσών τιμών με αντιγραφή παρατηρηθείσων τιμών από κάποιον δότη. Στην περίπτωση του συνολικού επιπέδου γνώσεων είναι επιθυμητό οι τιμές που θα προκύψουν από αντικατάσταση να είναι ακέραιες.

Όταν εφαρμόζεται en hot deck μέθοδος αντικατάστασης, οι τιμές που προκύπτουν από αντικατάσταση είναι όμοιες με τις παρατηρηθείσες. Γι αυτόν το λόγο είναι προτιμότερο στο στάδιο της αντικατάστασης, το επίπεδο γνώσεων να μοντελοποιηθεί κατ΄ αυτόν τον τρόπο και όχι με χρήση μοντέλου κανονικής παλινδρόμησης. Αντίθετα, στο πλαίσιο της ανάλυσης παλινδρόμησης, το σκορ μπορεί να μοντελοποιηθεί ως κανονική, εφόσον φυσικά ικανοποιούνται οι σχετικές υποθέσεις. Ο αριθμός των αντικαταστάσεων που έγιναν είναι 25. Τα μοντέλα για κάθε μεταβλητή παρουσιάζονται στο παράρτημα.

Οι αντικαταστάσεις έγιναν με τρεις διαφορετικούς τρόπους και στην συνέχεια τα ψευδο-πλήρη σετ δεδομένων αναλύθηκαν με τον ίδιο τρόπο. Το μόνο που διέφερε κατά την εφαρμογή των τριών διαφορετικών τρόπων ήταν ο ορισμός του μοντέλου αντικατάστασης. Στο πρώτο μοντέλο αντικατάστασης δεν εισάγονται τα βάρη, αγνοώντας έτσι τον σχεδιασμό της μελέτης. Το δεύτερο μοντέλο αντικατάστασης που χρησιμοποιήθηκε ήταν σταθμισμένο ενώ στο τρίτο τα βάρη χρησιμοποιήθηκαν ως ανεξάρτητη μεταβλητή με έναν απλό γραμμικό όρο. Τα βάρη τα οποία χρησιμοποιήθηκαν είναι τα τελικά βάρη (διορθωμένα για μη ανταπόκριση και για ηλικία-φύλο). Και στα τρία μοντέλα αντικατάστασης χρησιμοποιήθηκαν οι ίδιες ανεξάρτητες μεταβλητές εκτός από το τρίτο το οποίο περιείχε επιπλέον τα βάρη.

Στον πίνακα 5.10 παρουσιάζονται οι μέσες τιμές του συνολικού επιπέδου γνώσεων σε κάθε υποομάδα ηλικίας, φύλου, εκπαιδευτικού επιπέδου και εισοδήματος και συνολικά και στον πίνακα 5.11 παρουσιάζονται τα αποτελέσματα της πολλαπλής γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή το συνολικό επίπεδο γνώσεων ανάλογα με το μοντέλο αντικατάστασης που χρησιμοποιήθηκε.

Συνολικό επίπεδο γνώσεων	ΠΑ(α) Μέση Τιμή (95%ΔΕ)	ΠΑ(β) Μέση Τιμή (95%ΔΕ)	ΠΑ(γ) Μεσή Τιμή (95%ΔΕ)
Ηλικία			
< 40	0.341(-0.6, 1.28)	0.27(-0.667,1.22)	0.369(-0.53,1.27)
> 40	-3.03(-3.7,-2.35)	-3.029(-3.70,-2.35)	-2.99 (-3.668,-2.319)
Φύλο			
Άνδρες	-1.54(-2.31,-0.780)	-1.513(-2.25,-0.7765)	-1.512 (-2.27,-0.747)
Γυναίκες	-2.151 (-2.97,-1.33)	-2.17(-3.004, -1.336)	-2.115(-2.93,-1.29)
Επίπεδο εκπαίδευσης			
Μέχρι Δημοτικό	-6.033(-6.88,-5.178)	-6.096 (-6.89,-5.30)	-6.03(-6.83,-5.22)
Μέχρι Λύκειο	-1.85(-2.67,-1.027)	-1.82 (-2.63,-1.006)	-1.825(-2.62,-1.024)
Πάνω από Λύκειο	1.305(0.371,2.238)	1.339 (0.463,2.21)	1.36(0.487,2.25)
Εισόδημα			
< 900 €	-3.58(-4.51,-2.66)	-3.57(-4.49,-2.65)	-3.548(-4.45,-2.638)
900-1700€	-0.4146(-1.17,0.34)	-0.427(-1.2,0.346)	-0.418(-1.19,0.353)
> 1700 €	0.719(-1.11,2.55)	0.688(-1.11,2.49)	0.731(-1.069,2.53)
Συνολικά	-1.86(-2.46,-1.252)	-1.854(-2.45,-1.259)	-1.825(-2.42,-1.22)

Πίνακας 5.10: Μέση τιμή του συνολικού επιπέδου γνώσεων συνολικά και ανά κατηγορίες.
(α) Χωρίς χρήση βαρών στο μοντέλο αντικατάστασης
(β) Σταθμισμένο μοντέλο αντικατάστασης
(γ) Με τα βάρη ως γραμμικό όρο στο μοντέλο αντικατάστασης
ΔΕ: Διάστημα εμπιστοσύνης

Συγκρίνοντας τους συντελεστές παλινδρόμησης του πίνακα 5.11 με τους αντίστοιχους της τέταρτης στήλης του πίνακα 5.9 μπορούμε να εξάγουμε τα εξής συμπεράσματα σχετικά με το πως επη-

Μεταβλητή	ΠΑ(α) Συντελεστής 95% ΔΕ	ΠΑ(β) Συντελεστής 95% ΔΕ	ΠΑ(γ) Συντελεστής 95% ΔΕ
Ηλικία			
Ανά έτος:	-0.07** (-0.1,-0.04)	-0.0669** (-0.095,-0.038)	-0.071** (-0.101,-0.042)
Φύλο			
Άνδρας	0		
Γυναίκα	-0.0036 (-0.91,0.9)	-0.0447 (-0.91,0.821)	-0.0286 (-0.888,0.831)
Επίπεδο εκπαίδευσης			
Μέχρι Δημοτικό	0		
Μέχρι Λύκειο	2.753** (1.54,3.95)	2.85 ** (1.69,4.004)	2.777** (1.59,3.960)
Πάνω από Λύκειο	5.002** (3.4,6.6)	5.152** (3.622,6.681)	5.033 ** (3.47,6.58)
Εισόδημα			
<900 €	0		
900-1700€	1.68* (0.638,2.730)	1.69* (0.67,2.71)	1.67* (0.640,2.702)
>1700 €	1.869* (0.085,3.653)	1.931 (0.046,3.81)	1.85* (-0.0187,3.724)
Σταθερά	-1.96 (-4.38,0.4418)	-2.26 (-4.49,-0.0208)	-1.8833 (-4.22,0.457)

Πίνακας 5.11: Γραμμική παλινδρόμηση με εξαρτημένη μεταβλητή το συνολικό επίπεδο γνώσεων μετά από πολλαπλή αντικατάσταση . Σύγκριση των σημειακών εκτιμήσεων και των διαστημάτων εμπιστοσύνης.
(α) Χωρίς χρήση βαρών στο μοντέλο αντικατάστασης
(β) Σταθμισμένο μοντέλο αντικατάστασης
(γ) Με τα βάρη ως γραμμικό όρο στο μοντέλο αντικατάστασης
* p-value < 0.05
** p-value < 0.001

ρέασε τα αποτελέσματα η εφαρμογή πολλαπλών αντικαταστάσεων για μη ανταπόκριση ερώτησης:

- Το φύλο δεν συσχετίζεται με το επίπεδο γνώσεων σε καμία από τις αναλύσεις.
- Τα συμπεράσματα για την συσχέτιση της ηλικίας με το επίπεδο γνώσεων, επί τις ουσίας, δεν διαφοροποιούνται σημαντικά αν και ο συντελεστής είναι λίγο μειωμένος μετά την εφαρμογή πολλαπλών αντικαταστάσεων.
- Τα συμπεράσματα για την συσχέτιση της εκπαίδευσης με το επίπεδο γνώσεων επηρεάζονται περισσότερο. Κυρίως αλλάζει η διαφορά μεταξύ ατόμων μεταλυκειακού επιπέδου σε σύγκριση με άτομα επιπέδου μέχρι Λύκειο, η οποία αυξάνεται στις 2.5 μονάδες.
- Το εισόδημα είναι στατιστικά σημαντικό μετά από αντικαταστάσεις.

Τέλος, οι συντελεστές παλινδρόμησης δεν αλλάζουν ανάλογα με το μοντέλο αντικατάστασης. Όμως σύμφωνα με την θεωρία (James R. Carpenter 2013), το μοντέλο το οποίο περιλαμβάνει

τα βάρη ως ανεξάρτητη μεταβλητή δίνει τις πιο αμερόληπτες εκτιμήσεις, τόσο σημειακές όσο και διασποράς.

5.4 Χρήση προφυλακτικού

Ερωτήσεις οι οποίες σχετίζονται με προσωπικά θέματα εμφανίζουν συνήθως το μεγαλύτερο ποσοστό μη ανταπόκρισης. Μια τέτοια ερώτηση είναι η χρήση προφυλακτικού. Στόχος μας είναι να εκτιμήσουμε τα ποσοστά χρήσης προφυλακτικού συνολικά, αλλά και ανάλογα με την οικογενειακή κατάσταση. Εξαιρούμε από αυτή την ανάλυση άτομα που δήλωσαν ότι δεν έχουν ερωτικούς συντρόφους, καθώς και άτομα ηλικίας άνω των 70 ετών. Η ανάλυση γίνεται σε 414 άτομα τα χαρακτηριστικά των οποίων φαίνονται στον πίνακα 5.12. Συχνή χρήση προφυλακτικού γίνεται σε ποσοστό 22.9%, ενώ σε υψηλό ποσοστό (15.94%) η χρήση προφυλακτικού είναι άγνωστη. Τα ποσοστά διαφοροποιούνται όταν υπολογιστούν ξεχωριστά σε έγγαμους (ή σε συμβίωση) και μη έγγαμους. Παρόμοια υψηλά είναι τα ποσοστά των άγνωστων τιμών στις ερωτήσεις αριθμός ερωτικών συντρόφων και διάγνωση με Σεξουαλικώς Μεταδιδόμενο Νόσημα (ΣΜΝ).

5.4.1 Ανάλυση Πλήρων Παρατηρήσεων

Στην συνέχεια εκτιμούμε τα ποσοστά χρήσης προφυλακτικού συνολικά καθώς και σε έγγαμους και μη, με την χρήση των μεθόδων στάθμισης που εφαρμόστηκαν και προηγουμένως, αγνοώντας όμως την ύπαρξη ελλειπουσών τιμών. Τα αποτελέσματα παρουσιάζονται στον πίνακα 5.4.1. Φαίνεται ότι οι εκτιμήσεις των δύο πρώτων αναλύσεων υποεκτιμούν το ποσοστό χρήσης προφυλακτικού (29% και 31% αντίστοιχα). Τα αποτελέσματα της 3ης ανάλυσης είναι πολύ διαφορετικά (39% χρήση προφυλακτικού), καθώς υπάρχει συσχέτιση μεταξύ ηλικίας και χρήσης προφυλακτικού και οι νέοι είναι υποεκπροσωπημένοι στο δείγμα, είναι λογικό η διορθωμένη για ηλικία εκτίμηση να είναι υψηλότερη. Συχνή χρήση προφυλακτικού γίνεται από τους μη έγγαμους, ενώ αντίθετα η χρήση προφυλακτικού είναι λιγότερο συχνή στους έγγαμους.

Μεταβλητή	N (%)
Φυλο	
Άνδρας	190(45.89)
Γυναίκα	224(54.11)
Έγγαμος ή συμβίωση	
Ναι	280(67.63)
Όχι	134(32.37)
Χρήση Προφυλακτικού	
Ναι	95(22.95)
Όχι	253(61.11)
Άγνωστο	66(15.94)
Χρήση Προφυλακτικού(Έγγαμοι)	
Ναι	24(8.57)
Όχι	212(75.71)
Άγνωστο	44(15.71)
Σύνολο	280
Χρήση Προφυλακτικού(όχι έγγαμοι)	
Ναι	71(52.99)
Όχι	41(30.60)
Άγνωστο	22(16.42)
Σύνολο	134
Αριθμός συντρόφων	
1-5	248(59.90)
5+	107(25.85)
Άγνωστο	59(14.25)
Διάγνωση με ΣΜΝ	
Ναι	20(4.83)
Όχι	334(80.68)
Άγνωστο	60(14.49)

Πίνακας 5.12: Χαρακτηριστικά του υποσυνόλου του δείγματος (414 άτομα) που συμμετέχει στην ανάλυση για τη χρήση προφυλακτικού.

Μεταβλητή	Sample weights % (95%ΔΕ)	NR weights % (95%ΔΕ)	PS weights % (95%ΔΕ)
Χρήση Προφυλακτικού			
Ναι	28.27	31.91	39.7
	(22.55,34.79)	(26.9,37.38)	(32.74,47.12)
Όχι	71.73	68.09	60.3
	(65.21,77.45)	(62.62,73.1)	(52.88,67.26)
Χρήση Προφυλακτικού (Έγγαμοι)			
Ναι	10.57	14.27	15.82
	(6.93,15.8)	(9.64,20.61)	(10.47,23.21)
Όχι	89.43	85.73	84.18
	(84.2,93.07)	(79.39,90.35)	(76.79,89.53)
Χρήση Προφυλακτικού (όχι έγγαμοι)			
Ναι	63.94	66.9	75.19
	(53.82,72.96)	(56.61,75.8)	(65.47,82.89)
Όχι	36.06	33.1	24.81
	(27.04,46.18)	(24.2,43.39)	(17.11,34.53)

Πίνακας 5.13: Ανάλυση πλήρων παρατηρήσεων και με βάση τον σχεδιασμό.
Εκτίμηση του ποσοστού χρήσης προφυλακτικού με διαφορετικές μεθόδους στάθμισης.
Base weights: Ανάλυση με τα δειγματοληπτικά βάρη
NR weights: Ανάλυση με τα διορθωμένα για μη ανταπόκριση βάρη
PS weights : Ανάλυση με τα διορθωμένα βάρη για μη ανταπόκριση και για κατανομή ηλίκιας- φύλου.

Αυτές οι εκτιμήσεις δεν λαμβάνουν υπόψιν τις ελλείπουσες τιμές. Στην περίπτωσή μας τα ποσοστά των ελλειπουσών τιμών είναι ιδιαιτέρως υψηλά, γεγονός που πιθανώς να εισάγει μεροληψία στις εκτιμήσεις μας. Είναι λογικό να υποθέσουμε ότι οι ελλείπουσες τιμές δεν εμφανίζονται εντελώς τυχαία(MCAR). Υποθέτοντας ότι εμφανίζονται τυχαία (MAR) μπορούμε να εφαρμόσουμε κατάλληλες μεθόδους ώστε να λάβουμε υπόψιν τις άγνωστες τιμές. Δύο κατηγορίες μεθόδων θα εφαρμοστούν. Η πρώτη κατηγορία περιλαμβάνει τις μεθόδους στάθμισης με αντίστροφη πιθανότητα (ΣΑΠ), ενώ η δεύτερη τις πολλαπλές αντικαταστάσεις, όπως περιγράφτηκαν σε προηγούμενα κεφάλαια.

Στην συνέχεια θα γίνει σύγκριση των αποτελεσμάτων μεταξύ των μεθόδων. Από εδώ και πέρα θα γίνει χρήση μόνο των βαρών που χρησιμοποιήθηκαν για την 4η ανάλυση, δηλαδή των διορθωμένων βαρών για μη ανταπόκριση και για ηλικία-φύλο (τελικά βάρη).

5.4.2 Στάθμιση με Αντίστροφη Πιθανότητα(ΣΑΠ)

Θα εφαρμοστούν τρεις προσεγγίσεις αυτής της μεθόδου.

1. Η πρώτη προσέγγιση είναι η κλασική ΣΑΠ προσέγγιση. Με βάση αυτή εφαρμόζεται ένα μοντέλο λογιστικής παλινδρόμησης σε μια δείκτρια μεταβλητή :

$$R_i = \begin{cases} 1, & \text{αν το άτομο ανταποκρίθηκε} \\ 0, & \text{αλλιώς} \end{cases}$$

   Εδώ η ανταπόκριση έχει την έννοια ανταπόκρισης ερώτησης δηλαδή εάν το άτομο απάντησε στην ερώτηση. Αυτό το μοντέλο θα το αποκαλούμε μοντέλο ανταπόκρισης (ΜΑ), και από αυτό θα λάβουμε τις προβλεπόμενες πιθανότητες ανταπόκρισης. Στην συνέχεια θα χρησιμοποιήσουμε τις πιθανότητες αυτές για να επαναστάθμίσουμε τα τελικά βάρη. Έτσι χρησιμοποιώντας μόνο τους ανταποκριθέντες μπορούμε να λάβουμε διορθωμένες για μη ανταπόκριση ερώτησης εκτιμήσεις.

2. Η δεύτερη προσέγγιση έχει λογική αντικατάστασης. Εφαρμόζοντας μοντέλο λογιστικής παλινδρόμησης στην υπό μελέτη μεταβλητή, δηλαδή στη χρήση προφυλακτικού, μόνο στα άτομα με καταγεγραμμένη απάντηση, δοθέντων πλήρων επεξηγηματικών μεταβλητών μπορούμε να λάβουμε τις προβλεπόμενες πιθανότητες για όλα τα άτομα (ανταποκριθέντες και μη). Αυτό θα το αποκαλούμε μοντέλο έκβασης (ΜΕ). Η μέση τιμή των προβλεπόμενων πιθανοτήτων μπορεί να χρησιμοποιηθεί ως εκτίμηση της πιθανότητας χρήσης προφυλακτικού για όλο το δείγμα.

3. Η τρίτη προσέγγιση είναι η διπλά ανθεκτική (ΔΑ) εκτίμηση. Συνδυάζει τις δύο παραπάνω εκτιμήσεις ώστε η τελική εκτίμηση να είναι σωστή ακόμα και εάν ένα από τα μοντέλα ανταπόκρισης ή έκβασης δεν είναι σωστά προσδιορισμένο.

Οι εκτιμήσεις παρουσιάζονται στον πίνακα 5.14. Οι εκτιμήσεις που προκύπτουν από τις τρεις μεθόδους είναι παρόμοιες, με την χρήση προφυλακτικού να είναι στο 38% συνολικά, 14% για τους έγγαμους και 72% για τους άγαμους.

Μεταβλητή	ΜΑ % (95%ΔΕ)	ΜΕ % (95%ΔΕ)	ΔΑ % (95%ΔΕ)
Χρήση Προφυλακτικού			
Ναι	38.85	38.03	38.73
	(23.06,46.06)	(31.98,45.18)	(24.22,49.87)
Όχι	61.15	61.90	61.25
	(53.89,76.53)	(54.621 ,67.98)	(50.05,77.14)
Χρήση Προφυλακτικού (Έγγαμοι)			
Ναι	14.7	14.67	14.27
	(7.92,21.31)	(9.09,21.47)	(7.36,20.88)
Όχι	85.3	85.34	85.75
	(78.68,92.04)	(78.50,90.89)	(79.08,92.62)
Χρήση Προφυλακτικού (όχι έγγαμοι)			
Ναι	72.31	72.46	74.79
	(47.611,81.46)	(63.509,80.35)	(52.42,97.46)
Όχι	27.69	27.4	25.1
	(18.43,52.27)	(19.64,36.23)	(2.50,46.80)

Πίνακας 5.14: Ποσοστά χρήσης προφυλακτικού συνολικά και ανά οικογενειακή κατάσταση,με διόρθωση για μη ανταπόκριση. Χρήση μεθόδων στάθμισης με αντίστροφη πιθανότητα.
ΜΑ: Μοντέλο ανταπόκρισης
ΜΕ: Μοντέλο έκβασης
ΔΑ: Διπλά ανθεκτική
Τα ΔΕ υπολογίστηκαν με 1000 Bootstrap επαναλήψεις.

5.4.3 Ανάλυση μετά από Πολλαπλές αντικαταστάσεις

Η χρήση πολλαπλών αντικαταστάσεων (ΠΑ) είναι ιδιαίτερα διαδεδομένη τεχνική αντιμετώπισης μη ανταπόκρισης ερώτησης. Θα εφαρμόσουμε τέτοιες τεχνικές για να επανεκτιμήσουμε τα ποσοστά χρήσης προφυλακτικού. Οι μεταβλητές στις οποίες θα εφαρμοστούν αντικαταστάσεις είναι αυτές με μεγάλο αριθμό ελλειπουσών τιμών. Συγκεκριμένα η χρήση προφυλακτικού με 15.2% ελλείπουσες, ο αριθμός συντρόφων με 13.16% ελλείπουσες και η διάγνωση με κάποιο ΣΜΝ με 15.2% ελλείπουσες. Και οι τρεις μεταβλητές είναι δίτιμες επομένως μοντελοποιούνται με μοντέλο λογιστικής παλινδρόμησης, και οι αντικαταστάσεις θα γίνουν με χρήση αλυσιδωτών εξισώσεων.

Για την κάθε μεταβλητή διερευνήθηκαν πιθανοί προγνωστικοί παράγοντες με χρήση λογιστικής παλινδρόμησης. Οι ανεξάρτητες μεταβλητές με $p < 0.2$ χρησιμοποιήθηκαν στο μοντέλο αντικατάστασης. Τα μοντέλα για την κάθε μεταβλητή καθώς και τα διαγνωστικά για σύγκλιση και σταθερότητα του αλγορίθμου παρουσιάζονται στο παράρτημα. Η πρώτη ανάλυση ΠΑ έγινε αγνοώντας τα δειγματοληπτικά βάρη στο στάδιο των αντικαταστάσεων ενώ οι εκτιμήσεις που δίνονται είναι με βάση τον σχεδιασμό και χρήση των τελικών βαρών (με διόρθωση για μη ανταπόκριση και για την κατανομή ηλικίας-φύλου). Ο αριθμός Μ των αντικαταστάσεων είναι 25.

Στον πίνακα 5.15 παρουσιάζονται οι εκτιμήσεις οι οποίες προκύπτουν για το ποσοστό χρήσης προφυλακτικού. Θεωρώντας πιο αξιόπιστα τα αποτελέσματα που προκύπτουν με χρήση των τελικών βαρών ως γραμμικό όρο στο μοντέλο αντικατάστασης εξάγουμε τα εξής συμπεράσματα:

- Χρήση προφυλακτικού γίνεται σε ποσοστό 41.3%.
- Το ποσοστό αυτό διαφοροποιείται σημαντικά ανάλογα με την οικογενειακή κατάσταση.
- Η χρήση προφυλακτικού στους έγγαμους είναι 15.6%.
- Η χρήση προφυλακτικού στους άγαμους είναι 74.9%.

Τα αντίστοιχα συμπεράσματα που προκύπτουν από την εφαρμογή διπλά ανθεκτικής εκτίμησης είναι τα εξής:

- Χρήση προφυλακτικού γίνεται σε ποσοστό 38.7%.
- Το ποσοστό αυτό διαφοροποιείται σημαντικά ανάλογα με την οικογενειακή κατάσταση.
- Η χρήση προφυλακτικού στους έγγαμους είναι 14.2%.
- Η χρήση προφυλακτικού στους άγαμους είναι 74.79%.

Μεταβλητή	ΠΑ(α) % (95%ΔΕ)	ΠΑ(β) % (95%ΔΕ)	ΠΑ(γ) % (95%ΔΕ)
Χρήση Προφυλακτικού			
Ναι	41.52	41.54	41.37
	(32.49,50.54)	(32.5,50.57)	(32.31,50.42)
Όχι	58.47	58.45	58.62
	(49.45,67.50)	(49.42,67.49)	(49.57,67.68)
Χρήση Προφυλακτικού (Έγγαμοι)			
Ναι	15.44	15.39	15.68
	(9.7,21.6)	(9.76,21,03)	(9.88,21.48)
Όχι	84.55	84.60	84.31
	(78.83,90.26)	(78.96,90.23)	(78.51,90.11)
Χρήση Προφυλακτικού (όχι έγγαμοι)			
Ναι	75.61	75.72	74.95
	(64.59,86.63)	(64.69,86.74)	(63.68,86.22)
Όχι	24.38	24.27	25.04
	(13.36,35.4)	(13.25,35.3)	(13.77,36.31)

Πίνακας 5.15: Εκτιμήσεις του ποσοστού χρήσης προφυλακτικού με Πολλαπλές αντικαταστάσεις.
(α) Χωρίς χρήση βαρών στο μοντέλο αντικατάστασης
(β) Σταθμισμένο μοντέλο αντικατάστασης
(γ) Με τα βάρη ως γραμμικό όρο στο μοντέλο αντικατάστασης

Η εκτίμηση του ποσοστού χρήσης προφυλακτικού έγινε με πολλούς τρόπους. Το παρατηρούμενο ποσοστό στο δείγμα είναι 27.30%. Με εφαρμογή ανάλυσης πλήρων παρατηρήσεων με βάση τον σχεδιασμό αλλά χωρίς καμία διόρθωση για μη ανταπόκριση, το εκτιμώμενο ποσοστό είναι 28.27%. Διορθώνοντας τα βάρη για μη ανταπόκριση ατόμου το ποσοστό αυξάνεται, και ειδικά μετά την εκ των υστέρων στρωματοποίηση, το εκτιμώμενο ποσοστό είναι 39.7%. Αυτή η αλλαγή είναι λογική καθώς τα νεότερα σε ηλικία άτομα, τα οποία είναι πιο πιθανό να χρησιμοποιούν προφυλακτικό, λαμβάνουν μεγαλύτερα βάρη. Χωρίς διόρθωση για μη ανταπόκριση το εκτιμώμενο ποσοστό θα ήταν σαφώς υποεκτιμημένο. Επιπλέον η ύπαρξη ελλειπουσών τιμών δημιουργεί σφάλμα. Με εφαρμογή στάθμισης με το αντίστροφο της πιθανότητας ανταπόκρισης το εκτιμώμενο ποσοστό είναι 38.85% και με εφαρμογή διπλά ανθεκτικής εκτίμησης 38.73%. Τέλος, με ανάλυση μετά από πολλαπλές αντικαταστάσεις το αντίστοιχο ποσοστό είναι 41.37%.

Συζήτηση

Σκοπός της παρούσας διπλωματικής ήταν η αξιολόγηση μέσω εφαρμογής, μεθόδων ανάλυσης δεδομένων από σύνθετες πληθυσμιακές μελέτες, παρουσία μη ανταπόκρισης, όταν δηλαδή υπάρχουν ελλείποντα δεδομένα. Σε τέτοιου είδους μελέτες, ακόμα και με 100% ανταπόκριση, μόνο η εφαρμογή εξειδικευμένων τεχνικών στατιστικής ανάλυσης μας επιτρέπει την εξαγωγή αμερόληπτων εκτιμήσεων. Αυτού του είδους η ανάλυση καλείται ανάλυση με βάση τον σχεδιασμό (design-based analysis), και ένα βασικό στοιχείο της είναι η χρήση βαρών (weights). Η διαχείριση των βαρών πριν την στατιστική ανάλυση είναι ένα πολύ σημαντικό θέμα καθώς μπορεί να επηρεάσει άμεσα τόσο τις σημειακές εκτιμήσεις όσο και τις εκτιμήσεις διασποράς. Πριν την ανάλυση τα βάρη, συνήθως, τροποποιούνται ώστε να εμπεριέχουν διορθώσεις.

Στην παρούσα διπλωματική εφαρμόστηκε διόρθωση των βαρών για μη ανταπόκριση μέσω στάθμισης με το αντίστροφο του ποσοστού ανταπόκρισης ανά περιοχή. Η διόρθωση αυτή δεν βρέθηκε επαρκής. Εφαρμόστηκε επιπλέον διόρθωση με χρήση δεδομένων του πληθυσμού, γνωστή και ως εκ των υστέρων στρωματοποίηση. Αυτού του είδους η διόρθωση αποδείχθηκε σημαντική. Συγκεκριμένα, κατά την εφαρμογή ανάλυσης με βάρη τα οποία περιείχαν μόνο την πρώτη διόρθωση, το εκτιμώμενο ποσοστό για τους απασχολούμενους και τους συνταξιούχους ήταν 36.4% και 36.3% αντίστοιχα. Κατά την εφαρμογή ανάλυσης με βάρη τα οποία περιείχαν και την δεύτερη διόρθωση, το εκτιμώμενο ποσοστό για τους απασχολούμενους και τους συνταξιούχους ήταν 43.7% και 26.4% αντίστοιχα. Η αλλαγή στην συγκεκριμένη εκτίμηση είναι εντυπωσιακή αλλά προς την σωστή κατεύθυνση και αποδίδεται στην διαφορετική κατά ηλικία σύνθεση του δείγματος σε σύγκριση με την αντίστοιχη του πληθυσμού.

Με βάση τα παραπάνω, το κύριο συμπέρασμα είναι ότι η διόρθωση για μη ανταπόκριση μέσω στάθμισης με το αντίστροφο της πιθανότητας ανταπόκρισης δεν επαρκεί στην περίπτωση μη ανταπόκρισης ατόμου, ιδιαίτερα όταν η πληροφορία σχετικά με τους μη ανταποκριθέντες περιορίζεται στην περιοχή. Σε αυτές τις περιπτώσεις, η διόρθωση μέσω εκ των υστερών στρωματοποίησης κρίνεται απαραίτητη καθώς κάποιες βασικές μεταβλητές (π.χ. ηλικία και φύλο) σχετίζονται με πολλές από τις μεταβλητές ενδιαφέροντος.

Ακόμα μια τροποποίηση των βαρών, η οποία είναι σημαντική, αλλά δεν εφαρμόστηκε στα πλαίσια της παρούσας διπλωματικής είναι η περικοπή των βαρών (weight trimming) σε περίπτωση ακραία μεγάλων τιμών. Μετά από γραφική εποπτεία της κατανομής των βαρών βρέθηκε μια ακραία τιμή, όμως δεν κρίθηκε απαραίτητη η περικοπή της. Σε κάθε περίπτωση το θέμα παρουσιάζει ενδιαφέρον καθώς η περικοπή των βαρών είναι μια διαδικασία η οποία μπορεί να αλλάξει σημαντικά τα αποτελέσματα, αλλά μέχρι σήμερα γίνεται εντελώς αυθαίρετα (π.χ. στο 99ο ποσοστημόριο). Με το θέμα έχει ασχοληθεί ο Potter εκτενώς (Potter 1988, 1990). Άλλες, πιο σύγχρονες προσεγγίσεις έχουν προταθεί από τους Beaumont και Chowdhury (Beaumont 2008, Chowdhury et al. 2007), σύμφωνα όμως με τους Henry και Vallaint και αυτές οι προσεγγίσεις παρουσιάζουν προβλήματα (Henry & Valliant 2012).

Επίσης μια μελέτη εξέτασης υγείας περιλαμβάνει συνήθως δύο στάδια, συμπλήρωση ερωτηματολογίου και συμμετοχή σε εξετάσεις υγείας. Ενδέχεται κάποιο άτομο με συμπληρωμένο ερω-

τηματολόγιο να αρνηθεί συμμετοχή στις εξετάσεις. Κατι τέτοιο αποτελεί μη ανταπόκριση, με το πλεονέκτημα ότι για τα άτομα εκείνα υπάρχουν δεδομένα από την προηγουμένη φάση της μελέτης. Σε τέτοια περίπτωση, μπορεί να εφαρμοστεί στάθμιση με αντίστροφη πιθανότητα ανταπόκρισης, όπως περιγράφτηκε στην ενότητα 2.4.4, στους ανταποκριθέντες της δεύτερης φάσης. Η εφαρμογή τέτοιας διόρθωσης παρουσιάζει ενδιαφέρον, και προτείνεται να εφαρμοστεί κατά την ανάλυση δεδομένων από τις εξετάσεις υγείας της EMENO. Τέτοια διόρθωση εφαρμόζεται από την NHANES (Mirel LB 2013), όμως δεν είναι γνωστό το πως επηρεάζονται οι εκτιμήσεις διασποράς, όταν γίνεται χρήση εκτιμώμενης πιθανότητας, η οποία όμως στην συνέχεια αντιμετωπίζεται ως γνωστή.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η επαυξημένη στάθμιση με το αντίστροφο της πιθανότητας ανταπόκρισης ή διπλά ανθεκτική εκτιμήτρια. Η μέθοδος εφαρμόστηκε για μη ανταπόκριση ερώτησης στην εκτίμηση του ποσοστού χρήσης προφυλακτικού, ερώτηση με υψηλό ποσοστό ελλειπουσών τιμών. Δεν φάνηκε σημαντική τροποποίηση των αποτελεσμάτων σε σύγκριση με την ανάλυση πλήρων παρατηρήσεων με χρήση διορθωμένων βαρών. Αυτό μπορεί να οφείλεται σε κακό προσδιορισμό των μοντέλων που χρησιμοποιήθηκαν ιδίως του μοντέλου ανταπόκρισης, καθώς ήταν πιο δύσκολο να βρεθούν μεταβλητές οι οποίες να συσχετίζονται με την πιθανότητα ανταπόκρισης. Επιπλέον λόγω δυσκολίας στον υπολογισμό των διαστημάτων εμπιστοσύνης με αναλυτικό τρόπο, επιλέχθηκε Bootstrap μέθοδος, η οποία, όπως ήταν αναμενόμενο, δεν έδωσε ακριβή αποτελέσματα (Cao et al. 2009). Σύμφωνα με τους Tsiatis et. al. (Anastasios A. Tsiatis n.d.), υπάρχει τρόπος υπολογισμού της διακύμανσης. Οι ίδιοι συγγραφείς κάνουν προτάσεις για την ελαχιστοποίηση της διασποράς της εκτιμήτριας, με χρήση διαφορετικής μεθόδου εκτίμησης των παραμέτρων των μοντέλων έκβασης και ανταπόκρισης.

Τέλος έγινε εφαρμογή πολλαπλών αντικαταστάσεων με χρήση αλυσιδωτών εξισώσεων. Παρά το γεγονός ότι η εφαρμογή πολλαπλών αντικαταστάσεων σε πληθυσμιακές μελέτες δεν είναι καινούρια πρακτική (Khare et al. 1993), πολλοί αναλυτές δεν λαμβάνουν υπόψιν τον σχεδιασμό κατά το στάδιο των αντικαταστάσεων (Reiter et al. 2006). Αυτό όμως είναι λανθασμένο καθώς σύμφωνα με τον Carpenter (James R. Carpenter 2013) το κύριο πρόβλημα στην αιτιολόγηση των πολλαπλών αντικαταστάσεων στο πλαίσιο μιας πληθυσμιακής μελέτης είναι η έλλειψη μεθόδων για την κατασκευή κατάλληλων αντικαταστάσεων (proper imputations) (Rubin 1987). Είναι επίσης γνωστό ότι η εφαρμογή του τύπου υπολογισμού διακύμανσης πολλαπλών αντικαταστάσεων σε μια σύνθετη πληθυσμιακή μελέτη θα δώσει μεροληπτικές εκτιμήσεις αφού τα ψευτο-πλήρη σετ δεδομένων δεν προέρχονται από τον ακριβή δειγματοληπτικό μηχανισμό (Fay 1993, Kott 1995). Ακόμα, ο Carpenter δείχνει ότι ο υπολογισμός της διακύμανσης σύμφωνα με τον τύπο του Rubin δεν θα περιέχει μεροληψία εάν τα βάρη εισαχθούν στο μοντέλο αντικατάστασης. Επιπλέον σημειώνει ότι δεν αρκούν τα βάρη αλλά χρειάζεται να περιληφθούν και όλες οι αλληλεπιδράσεις τουλάχιστον για τις πλήρεις μεταβλητές. Φυσικά αυτό μπορεί να δημιουργήσει πρόβλημα στην εφαρμογή του μοντέλου, επομένως χρειάζεται να γίνει κάποια επιλογή μεταβλητών. Αυτή η μεθοδολογία μπορεί να χρησιμοποιηθεί και με χρήση των διορθωμένων βαρών, και οι εκτιμήσεις θα είναι αμερόληπτες δεδομένου ότι τα βάρη έχουν υπολογιστεί σωστά.

Η εφαρμογή των πολλαπλών αντικαταστάσεων έγινε καταρχάς αδρά, στη συνέχεια σταθμίζοντας το μοντέλο αντικατάστασης και τέλος με εισαγωγή των τελικών βαρών στο μοντέλο αντικατάστασης με γραμμικό τρόπο. Η μέθοδος φάνηκε να αποδίδει καλύτερα σε σύγκριση με την διπλά ανθεκτική εκτιμήτρια κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού. Με χρήση πολλαπλών αντικαταστάσεων η σημειακή εκτίμηση αυξήθηκε κατά περίπου 2% σε σχέση με την ανάλυση πλήρων παρατηρήσεων. Αν και μεταξύ των τριών προσεγγίσεων πολλαπλών αντικαταστάσεων δεν παρατηρήθηκαν σημαντικές διαφορές ούτε στις σημειακές ούτε στις εκτιμήσεις διασποράς, λόγω καλύτερης θεωρητικής τεκμηρίωσης συνιστάται η τρίτη. Όμως, λόγω του ότι δεν περιλήφθηκαν όροι αλληλεπίδρασης μεταξύ των βαρών και των υπόλοιπων επεξηγηματικών μεταβλητών, ενδέχεται να οι εκτιμήσεις αυτές να μην είναι οι καλύτερες δυνατές.

Τέλος, υπάρχει και μια νέα, πιο περίπλοκη προσέγγιση του προβλήματος, η οποία περιλαμβάνει

την εφαρμογή multilevel μοντέλων αντικατάστασης, με τα βάρη να ορίζουν το δεύτερο επίπεδο. Αυτή η προσέγγιση, έχει ενδιαφέρον, όμως εδώ, δεν διερευνήθηκε καθόλου.

Περίληψη

Όταν ο δειγματοληπτικός σχεδιασμός είναι σύνθετος η ανάλυση των δεδομένων πρέπει να γίνεται με ειδικές τεχνικές οι οποίες συμπεριλαμβάνουν τόσο τα δειγματοληπτικά βάρη όσο και τα στάδια της δειγματοληψίας. Παρουσία μη ανταπόκρισης ατόμου τα βάρη πρέπει να τροποποιούνται καταλλήλως ούτως ώστε το σταθμισμένο δείγμα να είναι αντιπροσωπευτικό του πληθυσμού. Για να γίνει τέτοιου είδους διόρθωση ο ερευνητής πρέπει να συλλέξει δεδομένα για τα άτομα που δεν ανταποκρίθηκαν. Με βάση τα διαθέσιμα δεδομένα δημιουργούνται κατηγορίες μέσα στις οποίες υπολογίζονται τα ποσοστά ανταπόκρισης. Με πολλαπλασιασμό των δειγματοληπτικών βαρών με το αντίστροφο του ποσοστού ανταπόκρισης επιτυγχάνεται η διόρθωση. Όσο καλύτερα ορισμένες είναι οι κατηγορίες αυτές τόσο πιο επιτυχημένη η διόρθωση. Επειδή τα δεδομένα αυτά συνήθως είναι περιορισμένα, η διόρθωση μπορεί να γίνει με χρήση δεδομένων για τον πληθυσμό αναφοράς. Έτσι προσαρμόζεται η σταθμισμένη κατανομή του δείγματος με τρόπο τέτοιο ώστε να συμφωνεί με γνωστές τιμές για τον πληθυσμό.

Ακόμα και μετά από τις διορθώσεις για μη ανταπόκριση ατόμου, παραμένει το πρόβλημα της μη ανταπόκρισης ερώτησης (ελλείπουσες τιμές). Εφαρμόστηκαν δύο κατηγορίες μεθόδων για την αντιμετώπιση του προβλήματος, πολλαπλές αντικαταστάσεις και στάθμιση με αντίστροφη πιθανότητα.

Για τις πολλαπλές αντικαταστάσεις μας ενδιέφερε να συμπεριλάβουμε τον δειγματοληπτικό σχεδιασμό. Αυτό έγινε με δύο τρόπους: με στάθμιση του μοντέλου αντικατάστασης και με χρήση των βαρών ως γραμμικό όρο σε αυτό. Επιπλέον θα μπορούσαν να έχουν χρησιμοποιηθεί τα βάρη ως παράγοντας στρωματοποίησης του μοντέλου, μετά δηλαδή από κατηγοριοποίηση των βαρών να γίνει αντικατάσταση ξεχωριστά σε κάθε κατηγορία. Τέλος θα μπορούσαν να έχουν χρησιμοποιηθεί μικτά μοντέλα αντικατάστασης (multilevel models).

Η μέθοδος στάθμισης με αντίστροφη πιθανότητα και η διπλά ανθεκτική εκτιμήτρια που χρησιμοποιήθηκαν, αν και εύκολα υλοποιήσιμες, δεν δίνουν ακριβή αποτελέσματα, εφόσον τα διαστήματα εμπιστοσύνης υπολογίστηκαν με επαναληπτικές διαδικασίες. Ενδεχομένως, εάν ο υπολογισμός της διακύμανσης γινόταν αναλυτικά, το πρόβλημα αυτό δεν θα υπήρχε ή θα ήταν μικρότερο.

Προτείνεται λοιπόν να γίνεται χρήση πολλαπλών αντικαταστάσεων με τα βάρη ως γραμμικό όρο, καθώς η μέθοδος αυτή δίνει ακριβή αποτελέσματα και είναι εύκολα εφαρμόσιμη, αφού τα στατιστικά πακέτα διαθέτουν έτοιμες ρουτίνες.

Abstract

When the sample design of a survey is complex, design-based analysis should be performed. This kind of analysis takes into account the stages of the complex design and the survey weights. In order to deal with unit nonresponse, the survey weights are properly adjusted so that the respondent weighted sample is representative of the population. To perform such a correction the researcher must collect data for non responders. Then, using the available information, weighting classes are created. Within these classes response probabilities are considered constant. The corresponding adjusted weight is the product of the inverse of the estimated response probability and the survey weight. Subsequent adjustments can be performed in order to conform the respondent sample distribution to distributions from an external source, such as population census.

The second type of non response we dealt with, is item non response. Two broad categories of methods are usually employed to address this issue. Inverse probability weighting and multiple imputation.

Standard multiple imputation methods can be modified to incorporate design features. We used the fully adjusted weights to incorporate design features into the imputation model first by weighting the imputation model, and second by including them as a predictor in the imputation model. Alternative approaches that we did not implement are stratified imputations and multilevel imputation models.

Inverse probability weighting method and doubly robust estimator are easily implemented using statistical software, and yielded similar estimates as multiple imputation. The main drawback, is that variance estimates and confidence intervals were computed using bootstrap methods, thus they were not as precise as the ones obtained using multiple imputation. Moreover, the structure of the missing data in a survey setting, do not allow the use of these methods.

To conclude, when handling item non response in a survey setting, a good practice is to perform multiple imputation using the fully adjusted weights as a covariate. This practice is both theoretically sound and easily implemented using statistical software such as Stata.

Παράρτημα

Εκτίμηση ποσοστού χρήσης προφυλακτικού Μοντέλα που χρησιμοποιήθηκαν και διαγνωστικά

Στους πίνακες 5.16, 5.17 παρουσιάζονται τα μοντέλα ανταπόκρισης και έκβασης που χρησιμοποιήθηκαν για τις εκτιμήσεις του ποσοστού χρήσης προφυλακτικού κατά την εφαρμογή της μεθόδου Στάθμισης με το αντίστροφο της πιθανότητας ανταπόκρισης. Αντίστοιχα στους πίνακες 5.18 5.19 5.20 παρουσιάζονται τα μοντέλα αντικατάστασης που χρησιμοποιήθηκαν για την ίδια εκτίμηση όταν εφαρμόστηκαν μέθοδοι πολλαπλής αντικατάστασης. Με τα γραφήματα 5.45.55.6 ελέγχεται η σύγκλιση του αλγόριθμου, και με τα γραφήματα 5.7 5.8 5.8 η σταθερότητα του, για τα τρία μοντέλα αντικατάστασης που εφαρμόστηκαν. Στα γραφήματα 5.4 5.55.6 δεν φαίνεται να υπάρχει κάποια τάση με τις επαναλήψεις(όπως συνεχής αύξηση) επομένως ο αλγόριθμος θεωρείται ότι έχει συγκλίνει. Στα γραφήματα 5.7 5.8 5.8, συγκρίνονται διαφορετικές αλυσίδες από τις αντικαταστάσεις (εδώ οι τρείς πρώτες). Μας ενδιαφέρει οι αλυσίδες να μην παρουσιάζουν πολύ διαφορετική συμπεριφορά μεταξύ των αντικατατσάσεων, αυτό θα ήταν ένδειξη αστάθειας. Δεν παρατηρείται κάτι τέτοιο.

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλλο			
Άνδρας*	1		
Γυναίκα	0.788	0.636	(0.294 ,2.114)
Ηλικία			
	0.984	0.346	(0.951,1.018)
Επίπεδο εκπαίδευσης			
Μέχρι Δημοτικό	1		
Μέχρι Λύκειο	1.755	0.427	(0.438,7.030)
Πάνω από Λύκειο	1.895	0.319	(0.539,6.662)
Έγγαμος ή σε Συμβίωση			
Ναι	1		
Όχι	0.780	0.584	(0.320,1.901)
Αριθμός συντρόφων			
1-5	1		
5+	3.915	0.081	(0.846,18.116)
Άγνωστο	0.022	<0.001	(0.007,0.071)
ΣΜΝ			
Όχι	1		
Ναι	1.513	0.459	(0.506,4.524)
Άγνωστο	0.590	0.365	(0.189,1.846)

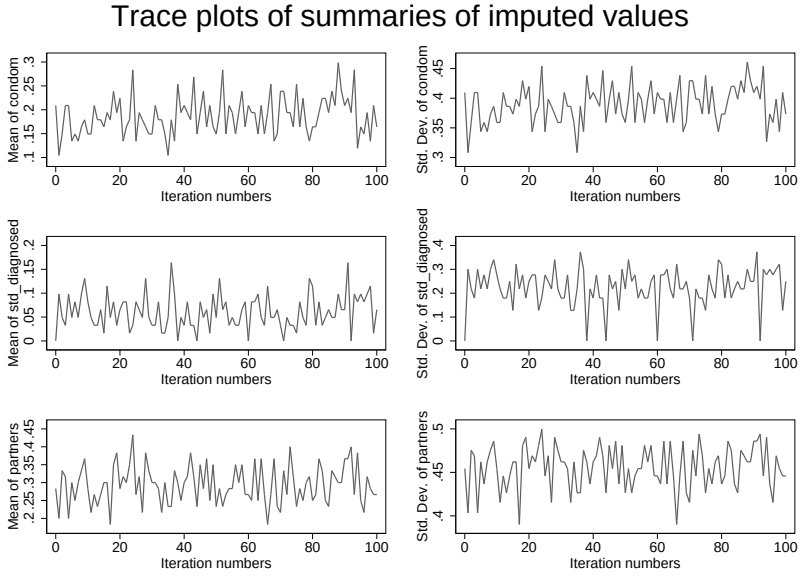
Πίνακας 5.16: Μοντέλο Λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή την ανταπόκριση στην ερώτηση χρήση προφυλακτικού. Οι προβλεπόμενες τιμές από το μοντέλο αυτό χρησιμοποιήθηκαν κατά την εφαρμογή της μεθόδου στάθμισης με το αντίστροφο της πιθανότητας ανταπόκρισης.

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλλο			
Άνδρας*	1		
Γυναίκα	1.388	0.540	(0.487,3.957)
Ηλικία			
Ανά έτος	0.947	0.069	(0.892,1.004)
Έγγαμος ή σε Συμβίωση			
Ναι	1		
Όχι	12.526	<0.001	(3.779,41.524)
Αριθμός συντρόφων			
1-5	1		
5+	3.242	0.003	(1.505,6.985)
Άγνωστο	1.340	0.738	(0.241,7.453)
ΣΜΝ			
Όχι	1		
Ναι	0.258	0.195	(0.033,2.004)
Άγνωστο	1.061	0.927	(0.302,3.728)
Έτη εκπαίδευσης			
Ανά έτος	1.401	<0.001	(1.221,1.608)
Αλκοόλ			
Καθόλου	1		
Μέτρια	1.306	0.678	(0.371,4.590)
Πολύ	2.210	0.080	(0.910,5.368)

Πίνακας 5.17: Μοντέλο λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή την χρήση προφυλακτικού. Οι προβλεπόμενες τιμές από το μοντέλο αυτό χρησιμοποιήθηκαν για την εκτίμηση του ποσοστού χρήση προφυλακτικού, με μονή αντικατάσταση (μοντέλο έκβασης).

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλλο			
Άνδρας*	1		
Γυναίκα	0.933	0.888	(0.358,2.432)
Ηλικία			
Ανά έτος	0.915	<0.001	(0.881,0.950)
Έγγαμος ή σε Συμβίωση			
Ναι	1		
Όχι	15.119	<0.001	(5.392,42.391)
Αριθμός συντρόφων			
1-5	1		
5+	4.2	0.007	(1.473,11.975)
ΣΜΝ			
Όχι	1		
Ναι	0.192	0.135	(0.022,1.675)
Έτη εκπαίδευσης			
Ανά έτος	1.555	<0.001	(1.318,1.836)

Πίνακας 5.18: Μοντέλο λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή τη χρήση προφυλακτικού. Κατά αυτόν τον τρόπο μοντελοποιήθηκε η μεταβλητή κατά την εφαρμογή πολλαπλών αντικαταστάσεων.



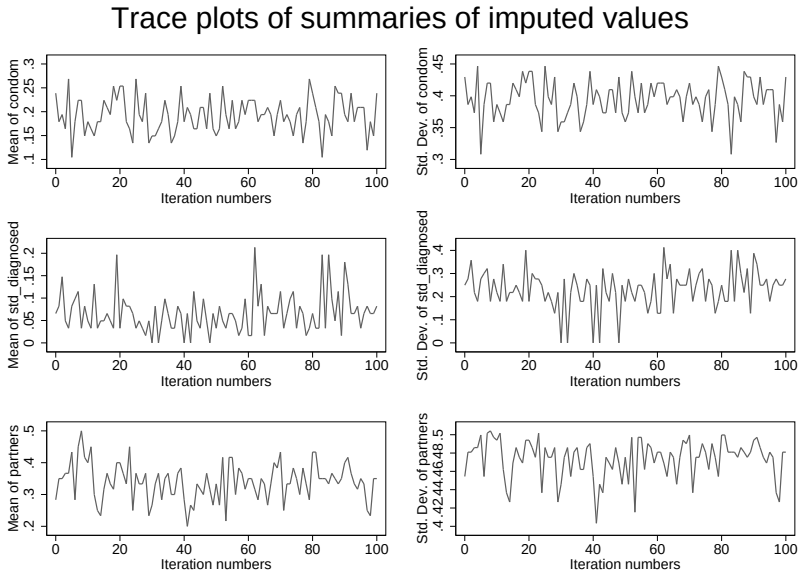
Σχήμα 5.4: Γράφημα για την σύγκλιση του MCMC αλγορίθμου κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης.

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλλο			
Άνδρας*	1		
Γυναίκα	1.577	0.478	(0.448,5.545)
Ηλικία			
Ανά έτος	0.964	0.186	(0.913,1.018)
Έγγαμος ή σε Συμβίωση			
Ναι	1		
Όχι	0.762	0.763	(0.131,4.434)
Αριθμός συντρόφων			
1-5	1		
5+	7.665	0.002	(2.057,28.565)
Χρήση Προφυλακτικού			
Όχι	1		
Ναι	0.437	0.409	(0.061,3.119)
Έτη εκπαίδευσης			
Ανά έτος	1.099	0.275	(0.927,1.303)

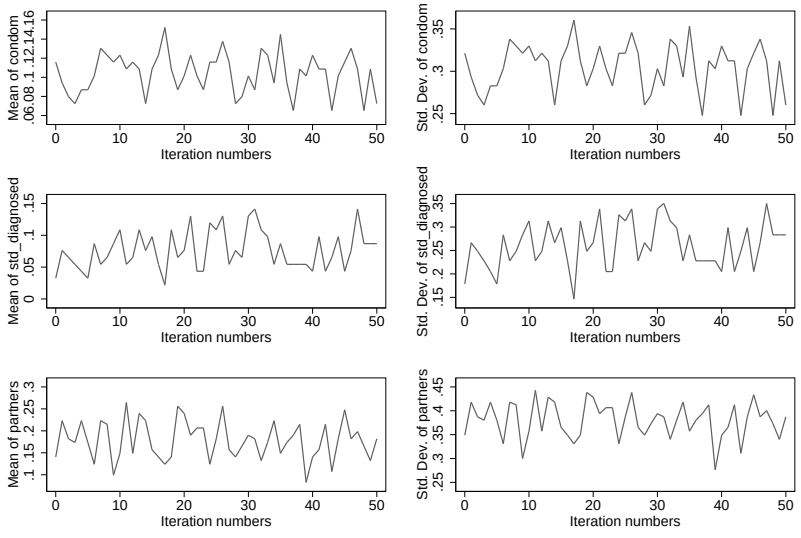
Πίνακας 5.19: Μοντέλο λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή την διάγνωση με ΣΜΝ. Κατά αυτόν τον τρόπο μοντελοποιήθηκε η μεταβλητή κατά την εφαρμογή πολλαπλών αντικαταστάσεων.

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλλο			
Άνδρας*	1		
Γυναίκα	0.384	0.012	(0.182,0.810)
Ηλικία			
Ανά έτος	1.001	0.922	(0.973 ,1.031)
Έγγαμος ή σε Συμβίωση			
Ναι	1		
Όχι	1.441	0.434	(0.578, 3.591)
Χρήση προφυλακτικού			
Όχι	1		
Ναι	2.409	0.073	(0.92,6.308)
Διάγνωση με ΣΜΝ			
Όχι	1		
Ναι	10.086	0.002	(2.412,42.173)
Εκπαίδευσης			
Μέχρι 6 έτη	1		
Μέχρι 12 έτη	13.084	0.002	(2.649,64.628)
Πάνω από 12 έτη	13.616	0.002	(2.585,71.712)
Αλκοόλ			
Καθόλου	1		
Μέτρια	1.459	0.384	(0.623,3.415)
Πολύ	2.724	0.034	(1.07,6.887)

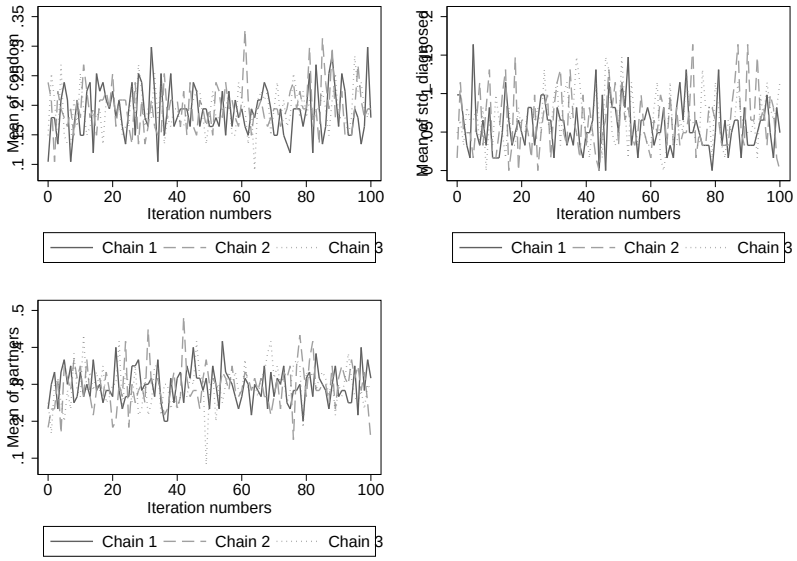
Πίνακας 5.20: Μοντέλο λογιστικής παλινδρόμησης για αριθμό ερωτικών συντρόφων. Κατά αυτόν τον τρόπο μοντελοποιήθηκε η μεταβλητή κατά την εφαρμογή πολλαπλών αντικαταστάσεων.



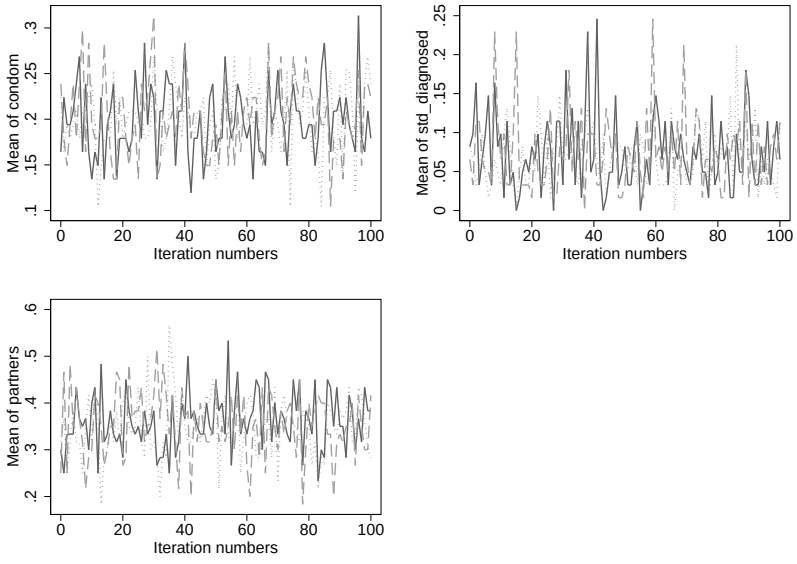
Σχήμα 5.5: Γράφημα για την σύγκλιση του MCMC αλγορίθμου κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού με πολλαπλή αντικατάσταση με χρήση σταθμισμένου μοντέλου αντικατάστασης.



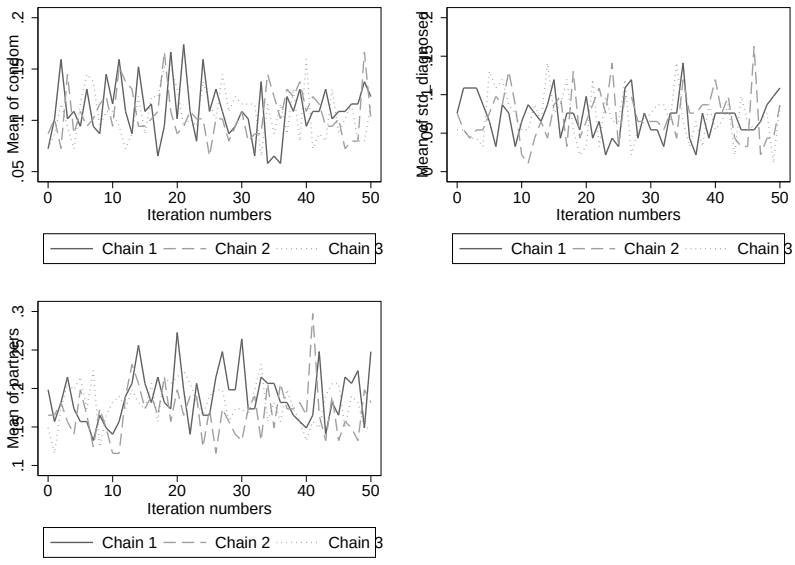
Σχήμα 5.6: Γράφημα για την σύγκλιση του MCMC αλγορίθμου κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης στο οποίο τα βάρη είχαν εισαχθεί ως γραμμικός όρος.



Σχήμα 5.7: Σύγκριση διαφορετικών αλυσίδων για την αξιολόγηση της σταθερότητας του MCMC αλγορίθμου κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης.



Σχήμα 5.8: Σύγκριση διαφορετικών αλυσίδων για την αξιολόγηση της σταθερότητας του MCMC αλγορίθμου κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού με πολλαπλή αντικατάσταση με χρήση σταθμισμένου μοντέλου αντικατάστασης.



Σχήμα 5.9: Σύγκριση διαφορετικών αλυσίδων για την αξιολόγηση της σταθερότητας του MCMC αλγορίθμου κατά την εκτίμηση του ποσοστού χρήσης προφυλακτικού με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης στο οποίο τα βάρη είχαν εισαχθεί ως γραμμικός όρος.

Επίπδο γνώσεων σχετικά με τον HIV
Μοντέλα αντικατάστασης που χρησιμοποιήθηκαν και διαγνωστικά

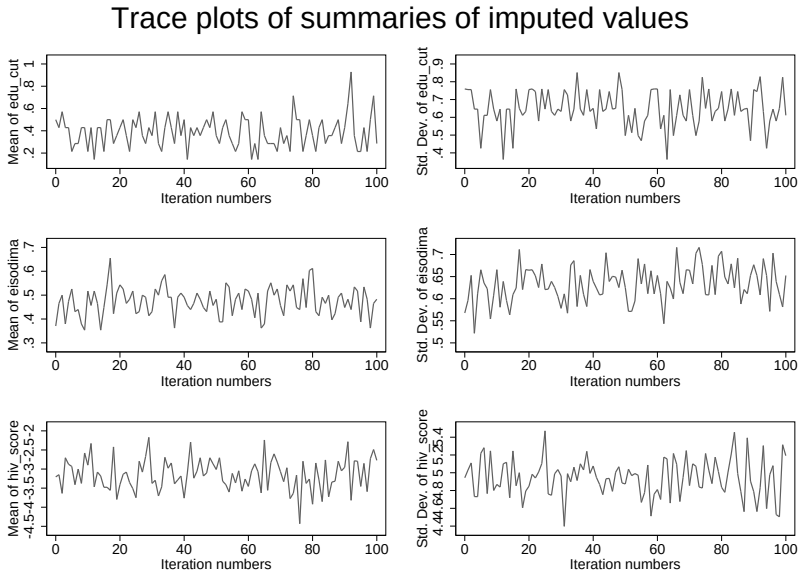
Στους πίνακες 5.21 5.22 παρουσιάζονται τα μοντέλα αντικατάστασης που χρησιμοποιήθηκαν στις ΠΑ για το συνολικό επίπεδο γνώσεων. Όπως προηγουμένως, με τα γραφήματα 5.10 5.11 5.12 ελέγχεται η σύγκλιση του αλγορίθμου και με τα γραφήματα 5.13 5.14 5.15 η σταθερότητα, για τις πολλαπλές αντικαταστάσεις ανάλογα με τον τύπο του μοντέλου αντικατάστασης που χρησιμοποιήθηκε. Με βάση τα γραφήματα δεν παρουσιάστηκαν προβλήματα σύγκλισης ή σταθερότητας στο στάδιο των αντικαταστάσεων.

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλο			
Άνδρα*	0		
Γυναίκα	-0.179	0.441	(-0.633,0.276)
Ηλικία			
ανά έτος	0.010	0.252	(-0.007,0.026)
Έτη εκπαίδευσης			
Μέχρι 6 έτη *	0		
Μέχρι 12 έτη	0.575	0.033	(0.048,1.103)
Πάνω από 12 έτη	1.752	<0.001	(1.111,2.392)
Άνεργος			
Όχι	0		
Ναι	-1.459	<0.01	(-2.192,-0.725)
Έγγαμος/Σε συμβίωση			
Ναι	0		
Όχι	-1.229	<0.001	(-1.769,-0.688)
Παιδιά			
Όχι	0		
Ναι	-0.408	0.206	(-1.041,0.224)
Κατανάλωση Αλκοόλ			
Καθόλου	0		
Μέτρια	0.877	<0.001	(0.390,1.365)
Πολύ	0.625	0.022	(0.090,1.161)
Συνολικό Σκορ για τον HIV			
	0.094	<0.001	(0.046,0.142)

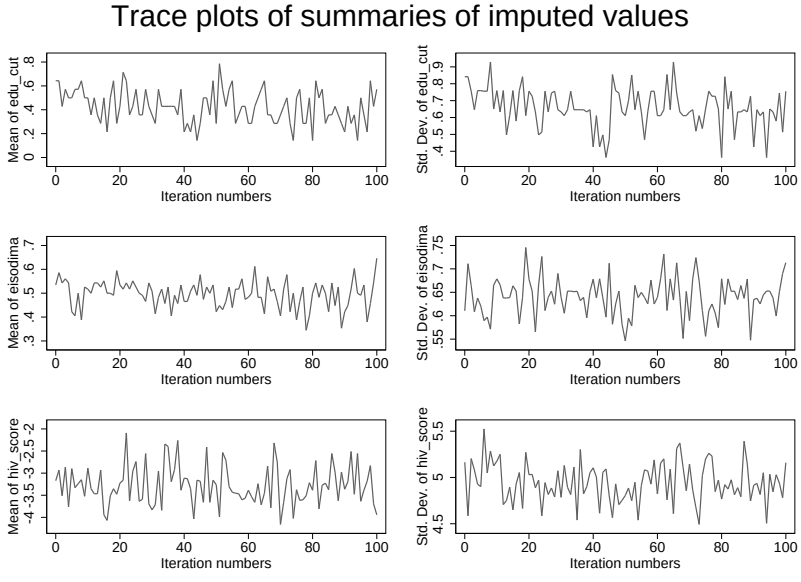
Πίνακας 5.21: Μοντέλο ordinal logistic με εξαρτημένη μεταβλητή το εισόδημα. Κατα αυτόν τον τρόπο μοντελοποιήθηκε η μεταβλητή κατά την εφαρμογή πολλαπλών αντικαταστάσεων.

Μεταβλητή	OR	p-value	95% ΔΕ
Φύλο			
Άνδρα*	0		
Γυναίκα	-0.150	0.457	(-0.547,0.246)
Ηλικία			
άνά έτος	-0.053	<0.001	(-0.068,-0.038)
Εισόδημα			
Μέχρι 900 *	0		
900-1700	0.748	0.001	(0.310,1.186)
Πάνω από 1700	1.841	<0.001	(1.087,2.596)
Άνεργος			
Όχι	0		
Ναι	-0.448	0.134	(-1.033,0.137)
Έγγαμος/Σε συμβίωση			
Ναι	0		
Όχι	0.418	0.099	(-0.079,0.914)
Παιδιά			
Όχι	0		
Ναι	-0.286	0.354	(-0.889,0.318)
Συνολικό Σκορ για τον HIV			
	0.179	<0.001	(0.133,0.224)

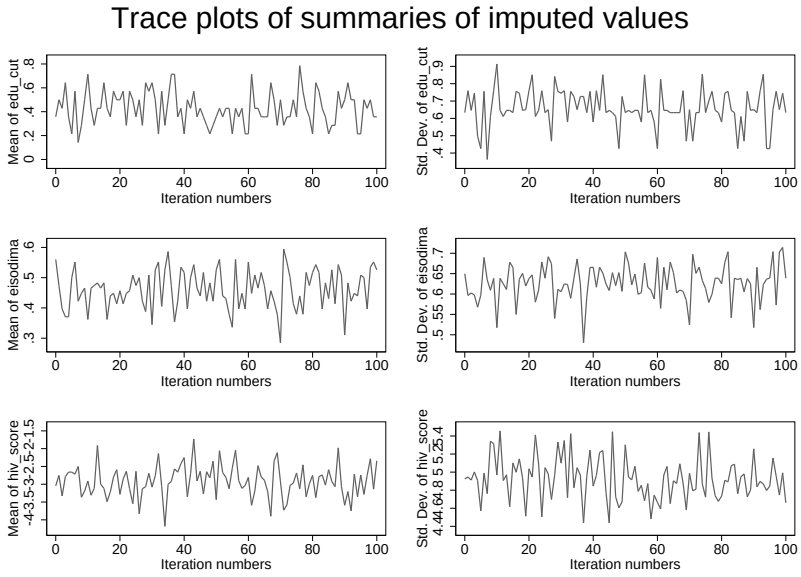
Πίνακας 5.22: Μοντέλο ordinal logistic με εξαρτημένη μεταβλητή το επίπεδο εκπαίδευσης. Κατα αυτόν τον τρόπο μοντελοποιήθηκε η μεταβλητή κατά την εφαρμογή πολλαπλών αντικαταστάσεων.



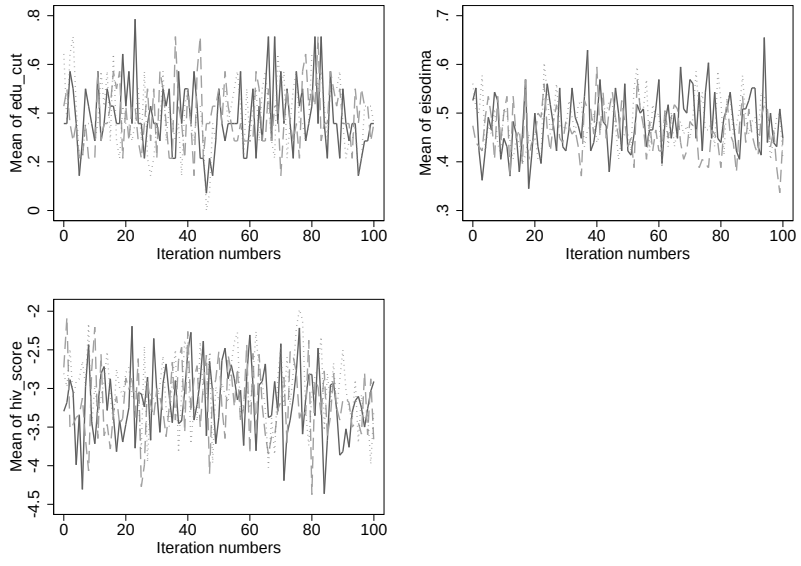
Σχήμα 5.10: Γράφημα για την σύγκλιση του MCMC αλγορίθμου κατά ανάλυση του συνολικού επιπέδου γνώσεων σχετικά με τον HIV με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης.



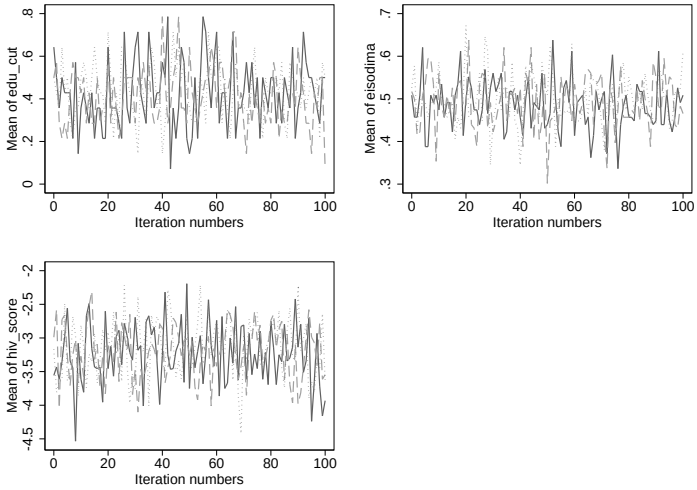
Σχήμα 5.11: Γράφημα για την σύγκλιση του MCMC αλγορίθμου κατά ανάλυση του συνολικού επιπέδου γνώσεων σχετικά με τον HIV με πολλαπλή αντικατάσταση με χρήση στάθσμένου μοντέλου αντικατάστασης.



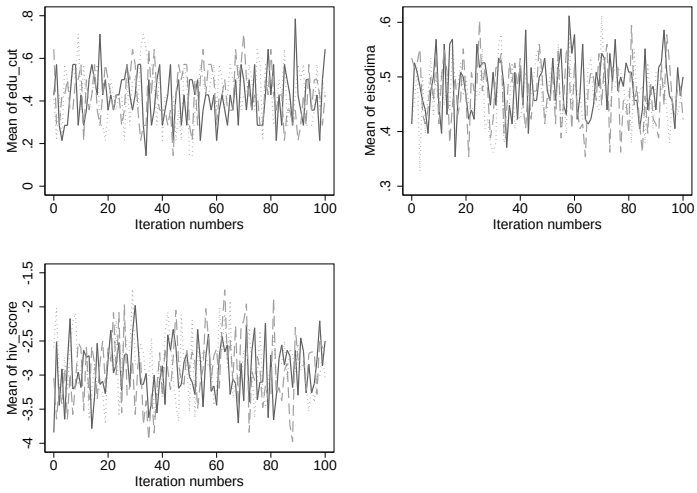
Σχήμα 5.12: Γράφημα για την σύγκλιση του MCMC αλγορίθμου κατά ανάλυση του συνολικού επιπέδου γνώσεων σχετικά με τον HIV με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης στο οποίο τα βάρη είχαν εισαχθεί ως γραμμικός όρος.



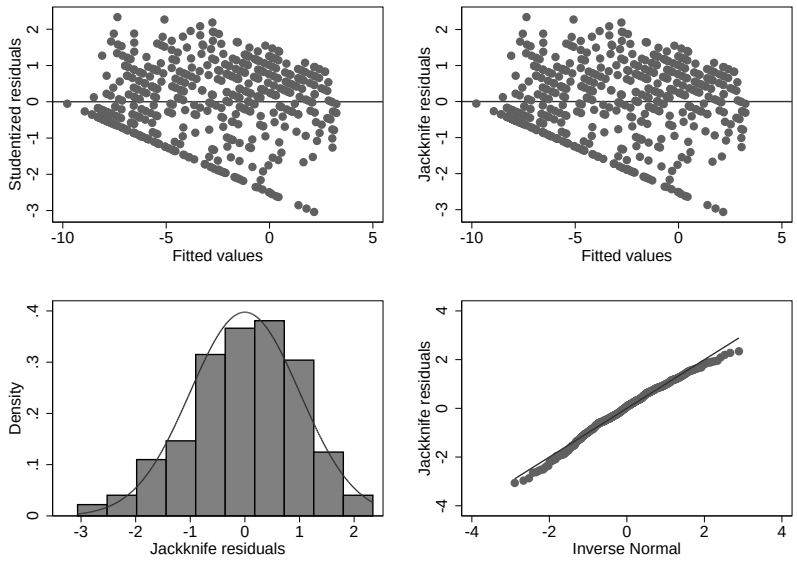
Σχήμα 5.13: Σύγκριση διαφορετικών αλυσίδων για την αξιολόγηση της σταθερότητας του MCMC αλγορίθμου κατά ανάλυση του συνολικού επιπέδου γνώσεων σχετικά με τον HIV με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης.



Σχήμα 5.14: Σύγκριση διαφορετικών αλυσίδων για την αξιολόγηση της σταθερότητας του MCMC αλγορίθμου κατά ανάλυση του συνολικού επιπέδου γνώσεων σχετικά με τον HIV με πολλαπλή αντικατάσταση με χρήση σταθμισμένου μοντέλου αντικατάστασης.



Σχήμα 5.15: Σύγκριση διαφορετικών αλυσίδων για την αξιολόγηση της σταθερότητας του MCMC αλγορίθμου κατά ανάλυση του συνολικού επιπέδου γνώσεων σχετικά με τον HIV με πολλαπλή αντικατάσταση με χρήση αστάθμιστου μοντέλου αντικατάστασης στο οποίο τα βάρη είχαν εισαχθεί ως γραμμικός όρος.



Σχήμα 5.16: Διαγνώστικά γραμμικού μοντέλου για το συνολικού επίπεδο γνώσεων.

Βιβλιογραφία

Anastasios A. Tsiatis, Marie Davidian, W. C. (n.d.), 'More robust doubly robust estimators'.
URL: *http://www4.stat.ncsu.edu/ davidian/doublerobust.pdf*

Announcement of the demographic and social characteristics of the Resident Population of Greece according to the 2011 Population - Housing Census (2013).
URL: *http://www.statistics.gr/portal/page/portal/ESYE/BUCKET/General/nws$_S$AM01$_E$N.PDF*

Arnold, B. C., Castillo, E. & Sarabia, J. M. (1999), *Conditional specification of statistical models*, Springer Science & Business Media.

Babbie, E. (2002), *The Practice of Social Research . 11th ed.*, Belmont, CA: Wadsworth.

Bang, H. & Robins, J. M. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**(4), 962–973.

Battaglia, M. P., Hoaglin, D. C. & Frankel, M. R. (2013), 'Practical considerations in raking survey data', *Survey Practice* **2**(5).

Beaumont, J.-F. (2008), 'A new approach to weighting and inference in sample surveys', *Biometrika* **95**(3), 539–553.

Bethlehem, J. (2002), *"Weighting Nonresponse Adjustments Based on Auxiliary Information.*

Bouhlila, D. S. & Sellaouti, F. (2013), 'Multiple imputation using chained equations for missing data in timss: a case study', *Large-scale Assessments in Education* **1**(1), 4.

Brick, J. M. (2013), 'Unit nonresponse and weighting adjustments: A critical review', *Journal of Official Statistics* **29**(3), 329–353.

Brick, J. M. & Kalton, G. (1996), 'Handling missing data in survey research', *Statistical methods in medical research* **5**(3), 215–238.

Cao, W., Tsiatis, A. A. & Davidian, M. (2009), 'Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data', *Biometrika* **96**(3), 723–734.

Carpenter, J. R. (2011), 'Multiple imputation with survey weights—a bad idea?'.

Celli BR, M. W. A. T. F. (2004), 'Standards for the diagnosis and treatment of patients with copd: a summary of the ats/ers position paper', *Eur Respir J, 23: 932-46* .

Chowdhury, S., Khare, M. & Wolter, K. (2007), Weight trimming in the national immunization survey, *in* 'Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association', pp. 2651–2658.

Cochran, W. G. (1997), *Sampling techniques*, John Wiley & Sons.

Collaborating Centre for Surveillance of Cardiovascular Diseases. Global Cardiovascular Infobase. (2010).
URL: *http://www.cvdinfobase.ca*

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–38.

Dillman, D. A., Eltinge, J. L., Groves, R. M. & Little, R. J. (2002), 'Survey nonresponse in design, data collection, and analysis', *Survey nonresponse* pp. 3–26.

Dowd, K., Kinsey, S., Wheeless, S. & Suresh, R. (2002), 'National survey of child and adolescent well-being (nscaw): Wave 1 data file user's manual', *Research Triangle Park, NC: Research Triangle Institute* .

Fay, R. E. (1993), 'Valid inferences from imputed survey data.', *Proceedings of the Survey Research Methodology Section of the American Statistical Association* pp. 41–48.

Gary, P. R. (2007), Adjusting for nonresponse in surveys, *in* 'Higher education: Handbook of theory and research', Springer, pp. 411–449.

Gelfand, A. E. & Smith, A. F. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American statistical association* **85**(410), 398–409.

Groves, R. M. (2006), 'Nonresponse rates and nonrespnse bias in household surveys', *Public Opinion Quarterly* **70**(5), 646–675.

Groves, R. M., . C. M. P. (1998), *Nonresponse in household interview surveys*, New York: Wiley.

Hansen, M. H. & Hurwitz, W. N. (1946), 'The problem of non-response in sample surveys', *Journal of the American Statistical Association* **41**(236), 517–529.

Henry, K. & Valliant, R. (2012), Methods for adjusting survey weights when estimating a total, *in* 'Proceedings of the 2012 Federal Committee on Statistical Methodology's Research Conference'.

Hinkley, D. (1985), 'Transformation diagnostics for linear models', *Biometrika* **72**(3), 487–496.

Hosmer Jr, D. W. & Lemeshow, S. (2004), *Applied logistic regression*, John Wiley & Sons.

James R. Carpenter, M. G. K. (2013), *Multiple imputation and its application - 1st ed*, John Wiley Sons, Ltd.

Jelke Bethlehem, B. S. (2004), 'Nonresponse adjustment in household surveys', *Statistics Netherlands* .

Jenkins, P., Earle-Richardson, G., Burdick, P. & May, J. (2008), 'Handling nonresponse in surveys: analytic corrections compared with converting nonresponders', *American journal of epidemiology* **167**(3), 369–374.

Jones, J. (1995), 'An illustration of the danger of nonrespnse for survey research. paper presented at the annual conference of the american educational research association,san francisco,ca.'.

Jones, J. (1996), 'The effects of non-response on statistical inference', *Journal of Health Social Policy* **8**(1), 49–62.

Kalton, G., FloresCervantes, I., Zheng, H., Little, R. J., Wu, C., Luan, Y., Lu, H., Gelman, A., de Leeuw, E. D., Hox, J. et al. (1998), 'Weighting methods', *New Methods for Survey Research* pp. 79–94.

Kalton, G. & Maligalig, D. S. (1991), A comparison of methods of weighting adjustment for nonresponse, *in* 'Proceedings of the US Bureau of the Census 1991 annual research conference', pp. 409–428.

Kang, J. D. & Schafer, J. L. (2007), 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data', *Statistical science* pp. 523–539.

Khare, M., Little, R. J., Rubin, D. B. & Schafer, J. L. (1993), Multiple imputation of nhanes iii, *in* 'Proceedings of the American Statistical Association, Survey Research Methods Section', p. 00.

Kim, J. K. & Haziza, D. (2010), Doubly robust inference with missing data in survey sampling, *in* 'Joint Statistical Meetings, Vancouver, Canada'.

Kim, J. K., Michael Brick, J., Fuller, W. A. & Kalton, G. (2006), 'On the bias of the multiple-imputation variance estimator in survey sampling', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3), 509–521.

Kott, P. (1995), 'A paradox of multiple imputation.', *Proceedings of the Survey Research Methodology Section of the American Statistical Association* pp. 380–383.

Lessler, J. & Kalsbeek, W. (1992), *Nonsampling Error in Surveys.*, New York: Wiley.

Liu, C. (2004), 'Robit regression: a simple robust alternative to logistic and probit regression', *Applied Bayesian modeling and causal inference from incomplete-data perspectives* pp. 227–238.

Mirel LB, Mohadjer LK, D. S. e. a. (2013), 'National health and nutrition examination survey: Estimation procedures, 2007–2010', *Vital Health Stat 2(159)* .

Panagiotakos, D. B., Chrysohoou, C., Pitsavos, C., Tzioumis, K., Papaioannou, I., Stefanadis, C. & Toutouzas, P. (2002), 'The association of mediterranean diet with lower risk of acute coronary syndromes in hypertensive subjects', *International journal of cardiology* **82**(2), 141–147.

Paul S.Levy, S. L. (2008), *Sampling of Populations. Methods and Applications 4th ed.*, Wiley.

Potter, F. (1988), 'Survey of procedures to control extreme sampling weights', *Proceedings of the Section on Survey Research Methods* pp. 453–458.

Potter, F. (1990), 'Study of procedures to identify and trim extreme sample weights', *Proceedings of the Section on Survey Research Methods* pp. 225–230.

Rabe KF, Hurd S, A. A. e. a. (2007), 'Global strategy for the diagnosis, management, and prevention of copd-2006 update.', *Am J Respir Crit Care Med* **176**.

Rebecca R. Andridge, R. J. A. L. (2010), 'A review of hot deck imputation for survey non-response', *International Statistical Review* **78**(1), 40–64.

Reiter, J. P., Raghunathan, T. E. & Kinney, S. K. (2006), 'The importance of modeling the sampling design in multiple imputation for missing data', *Survey Methodology* **32**(2), 143.

Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys. New York: J*, Wiley & Sons.

Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999), 'Adjusting for nonignorable drop-out using semiparametric nonresponse models(with discussion and rejoinder)', *Journal of the American Statistical Association* **94**(448), 1096–146.

Scheaffer RL, Mendenhall W, O. L. (1996), *Elementary Survey Sampling, 5th edition*, Duxbury Press,Boston.

Schenker, N. & Taylor, J. M. (1996), 'Partially parametric techniques for multiple imputation', *Computational Statistics & Data Analysis* **22**(4), 425–446.

Seaman, S. R. & White, I. R. (2013), 'Review of inverse probability weighting for dealing with missing data', *Statistical Methods in Medical Research* **22**(3), 278–295.

Seaman, S. R., White, I. R., Copas, A. J. & Li, L. (2012), 'Combining multiple imputation and inverse-probability weighting', *Biometrics* **68**(1), 129–137.

Singleton Jr, R. A. & Bruce, C. (n.d.), 'Straits. 2005', *Approaches to social research* **4**.

StataCorp (2013), *Stata 13 Multiple Imputation-Reference Manual*, College Station, TX: Stata Press.

Καλπινέλλη Ευαγγελία (2004), 'Το πρόβλημα των ελλειπόντων δεδομένων και οι νέοι τρόποι αντιμετώπισής του', *Πρακτική άσκηση,Οικονομικό Πανεπιστήμιο Αθηνών-Τμήμα Στατιστικής* . **URL:** *http://stat-athens.aueb.gr/ jpan/short-course-ergasia-Kalpinelli.pdf*

Tolonen, H., Koponen, P., Aromaa, A., Conti, S., Graff-Iversen, S., Grøtvedt, L., Kanieff, M., Mindell, J. et al. (2008), 'Recommendations for the health examination surveys in europe'.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. & Rubin, D. B. (2006), 'Fully conditional specification in multivariate imputation', *Journal of statistical computation and simulation* **76**(12), 1049–1064.

Wang, C., Wang, S., Zhao, L.-P. & Ou, S.-T. (1997), 'Weighted semiparametric estimation in regression analysis with missing covariate data', *Journal of the American Statistical Association* **92**(438), 512–525.

White, I. R., Daniel, R. & Royston, P. (2010), 'Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables', *Computational statistics & data analysis* **54**(10), 2267–2275.

Wirth, K. E., Tchetgen, E. J. T. & Murray, M. (2010), 'Adjustment for missing data in complex surveys using doubly robust estimation: Application to commercial sexual contact among indian men', *Epidemiology* **21**(6), 863–871.